# Machine learning en finance : vers de nouvelles stratégies ?

Vincent Bouchet



Mémoire de stage/ de recherche

# Machine Learning en Finance

## Vers de nouvelles stratégies ?


**Présenté par : BOUCHET Vincent**

**Nom de l'entreprise : Exane**
**Tuteur entreprise : ZIADE Jean-Baptiste**
**Tuteur universitaire : GIRERD-POTIN Isabelle**


**Master 2 Pro. FC**
**Master Finance d'Entreprise et Gestion des Risques**
**2016 - 2017**

# Machine Learning en Finance

*Vers de nouvelles stratégies ?*

Vincent BOUCHET - Tuteur : Isabelle GIRERD-POTIN

Juin 2017

Mémoire de Master 2 FEGR

Avertissement :

Grenoble IAE, au sein de l'Université Grenoble Alpes, n'entend donner aucune approbation ni improbation aux opinions émises dans les mémoires des candidats aux masters en alternance : ces opinions doivent être considérées comme propres à leur auteur.

Tenant compte de la confidentialité des informations ayant trait à telle ou telle entreprise, une éventuelle diffusion relève de la seule responsabilité de l'auteur et ne peut être faite sans son accord.

Résumé

Le Machine Learning connait un vif regain d'intérêt ces dernières années grâce au Big Data. En finance, les applications sont variées : tarification des assurances, gestion de crédits, détection des fraudes et gestion de portefeuille. Ces algorithmes ont entre autre permis de faire évoluer la recherche sur la question de la prédictibilité des cours. Les réseaux de neurones ainsi que les algorithmes hybrides sont les plus performants dans ce domaine, mais les précisions ne progressent pas pour autant (de l'ordre de 65%). Cela s'explique en partie par l'assimilation des différents acteurs des stratégies publiées par la recherche académique, et donc par la disparition progressive de ces anomalies. L'évolution de ces dernières années concerne principalement le type de données traitées : l'analyse de texte permet d'extraire les sentiments des réseaux sociaux. Les résultats sont cependant mitigés et confirment la nécessité d'apporter un regard métier sur l'analyse prédictive afin par exemple de développer des dictionnaires spécifiques au monde de la finance.

## DECLARATION ANTI-PLAGIAT

Ce travail est le fruit d'un travail personnel et constitue un document original. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.

Je m'engage sur l'honneur à signaler, dans le présent mémoire, et selon les règles habituelles de citation des sources utilisées, les emprunts effectués à la littérature existante et à ne commettre ainsi aucun plagiat.

**NOM, PRENOM**

Bouchet Vincent.

**DATE, SIGNATURE**

30/06/2017

# Remerciements

J’adresse mes remerciements aux personnes qui m’ont aidé dans la réalisation de ce mémoire.

En premier lieu, je remercie Mme Isabelle GIRERD-POTIN, professeur à l’IAE de Grenoble. En tant que directrice de mémoire, elle m’a guidé dans ce travail en me proposant des pistes de réflexion ainsi que des références pertinentes afin de construire et de structurer ce travail.

Je remercie aussi Jean-Baptiste ZIADE, responsable du Middle-Office Dérivés OTC chez Exane Derivatives. En tant que maitre de stage, il m’a énormément apporté à la fois en termes de connaissances sur les produits dérivés mais aussi sur la manière d’interagir et de m’exprimer avec les différents partenaires d’un Middle Office.

Je remercie toute l’équipe pédagogique de l’IAE de Grenoble et plus particulièrement du Master 2 FEGR pour cette année particulièrement riche en apprentissage et en découverte.

Je remercie également mes collègues, mes frères Florent et Martin pour les discussions que nous avons pu avoir autour de ce sujet, Alexia pour son soutien, et enfin mes parents pour la dernière relecture !

# Sommaire

# Introduction

Le Machine Learning (ou apprentissage automatique), branche de l'intelligence artificielle, connait un vif regain d'intérêt ces dernières années grâce au Big Bata (Gandomi and Haider, 2015). On considère que 90 % des données actuelles ont été générées sur les deux dernières années. A titre d'exemple, 5900 Tweets sont envoyés chaque seconde sur le réseau Tweeter[1], 350 millions de photos sont envoyées sur Facebook chaque jour.[2] Appliquée dans des domaines variés comme le marketing digital, la biologie, l'économie ou l'industrie automobile, l'analyse prédictive représente un enjeu stratégique considérable.

En finance, les applications sont multiples : tarification, détection de fraudes, gestion de portefeuilles. Elles concernent à la fois les banques de détails et les marchés financiers. Dans ce domaine, les données ont toujours été utilisées, mais la tendance était jusqu'à lors orientée vers la conception de modèles probabilistes complexes (processus stochastiques à temps discret ou continu). Le Machine Learning ouvre un champ de recherche complémentaire, qui se recentre sur l'exploitation des données (Nardo et al. 2016). Depuis l'hypothèse d'efficience des marchés (Fama, 1970) , la question de la prédictibilité des cours a été régulièrement étudiée, en particulier au travers d'études d'évènements, à la fois avec des méthodes traditionnelles (régressions, analyse en composante principale, analyse factorielle) et plus récemment grâce aux nouveaux algorithmes de Machine Learning.

L'objectif de ce travail de recherche est double : réaliser une revue des applications du Machine Learning en Finance, avant d'étudier le rôle des nouveaux algorithmes dans l'étude de la prédictibilité des cours et des stratégies de trading.

La première partie de ce mémoire vise à mieux cerner le champ d'action du Machine Learning ainsi que le fonctionnement des principaux algorithmes : n'étant pas issu d'une formation «informatique », il était essentiel de me familiariser avec ces concepts. J'ai ensuite étudié les applications du Machine Learning dans les domaines de l'assurance, de la gestion de crédit, de la détection de fraudes et de la gestion de portefeuille. La troisième partie s'appuie sur 42 études académiques publiées entre 1988 et 2017 pour comprendre le rôle des nouvelles données et algorithmes sur la question de la prédictibilité des cours.

[1] https://www.planetoscope.com/Internet-/1547-nombre-de-tweets-expedies-sur-twitter.html
[2] http://www.blogdumoderateur.com/chiffres-facebook/

# 1 Principes et évolution du Machine Learning

## 1.1 Histoire du Machine Learning

Le Machine Learning (ML) est né dans les années 80, et peut être considéré comme une branche de l'intelligence artificielle (IA, dont les débuts remontent à l'après guerre). Le ML (ou apprentissage automatique) est intimement lié à l'analyse de données et aux algorithmes de décision.

L'analyse de données tire son origine des statistiques. Jusqu'au 18eme siècle, la statistique, alors « science de l'état » était uniquement descriptive (Desrosières, 2010). Ce n'est qu'un siècle plus tard que les probabilités seront liées aux statistiques, avec entre autre la notion d'extrapolation entre l'observation d'un échantillon et les caractéristiques d'une population. A partir du début du 20eme siècle, les statistiques s'organisent comme une science à part et deux disciplines se distinguent : les statistiques descriptives et les statistiques inférentielles[3].

La notion d'apprentissage automatique fait d'abord référence à la compréhension de la pensée humaine, étudiée par Descartes puis par G. Leibniz dans son ouvrage « De Arte combinatoria » en 1666. Le philosophe tente alors de définir les raisonnements les plus simples de la pensée (à l'image d'un alphabet) qui permettront, une fois combinés de formuler des pensées très complexes[4].Ces travaux seront formalisés par G.Boole en 1854 dans son ouvrage « *An investigation of the laws of thought, on which are founded the mathematical of logic and Probability* ».

L'IA prend un tournant en 1956 avec la conférence de Dartmouth organisée par M. Minsky et J. McCarthy. Durant cet évènement, un principe proche de celui de Leibniz est formulé : « *chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut être si précisément décrit qu'une machine peut être conçue pour le simuler* »[5]. Il faudra cependant attendre les années 80 et le développement des capacités de calcul pour que l'IA se

[3] https://fr.wikipedia.org/w/index.php?title=Histoire_de_l%27intelligence_artificielle&oldid=132868931
[4] https://en.wikipedia.org/w/index.php?title=De_Arte_Combinatoria&oldid=782514244
[5] http://blog.octo.com/une-histoire-de-la-data-science-par-deux-data-scientists/

développe. Les lois de Moore, qui conjecturent l'évolution de la puissance des ordinateurs, prévoient une multiplication par deux des capacités de calcul chaque année[6].

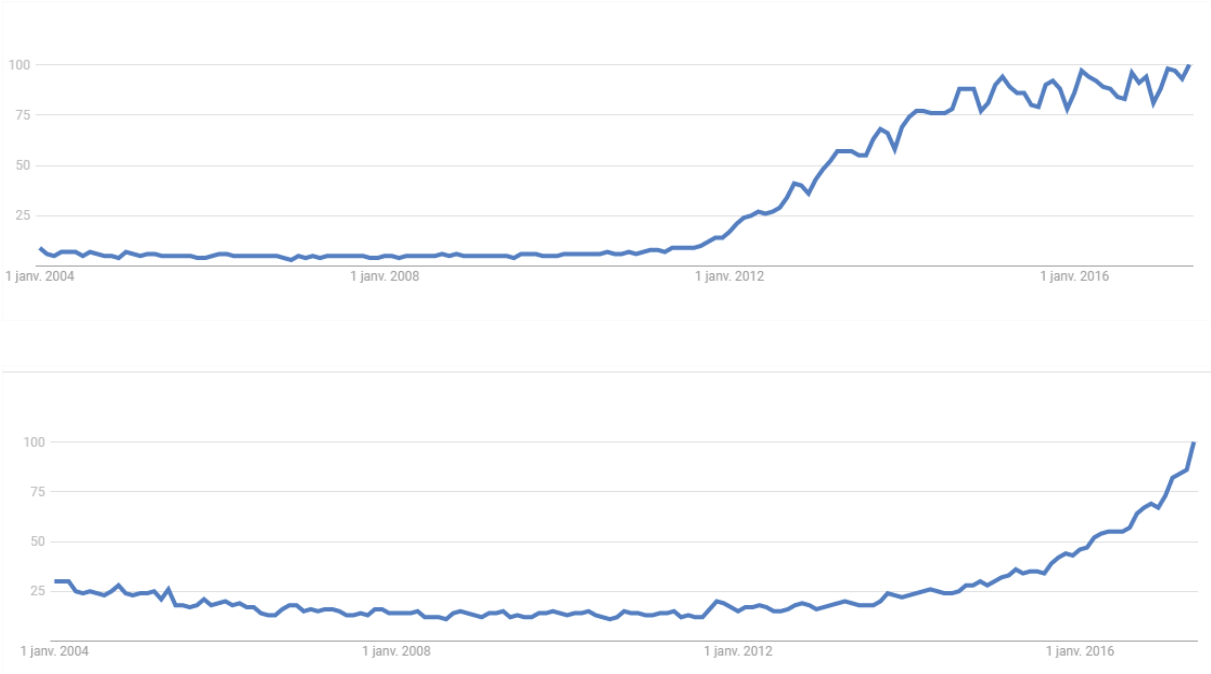
En 1982, J.Hopfield développe un nouveau type de réseau neuronal (réseau de Hopfield), qui sera largement repris par la suite[7]. On peut aussi retenir 1984 comme une date clef dans la naissance du ML avec l'introduction des arbres de décision par L. Breiman, J.Friedman, R. Olshen et C. Stones En 2001, L. Breiman et A.Cutler étendent ce concept avec les forêts aléatoires (« random forests »)[8], un modèle que nous étudierons par la suite.

En termes d'applications marquantes du Machine Learning, l'ordinateur IBM Deep Blue bat le champion du monde d'échec Garry Kasparov[9] en 1997, mais si l'on peut parler d'intelligence artificielle, il ne s'agit pas d' « apprentissage automatique » à proprement parlé. En 2014, c'est un chatbot russe « Eugene Goostman » qui devient le premier « robot » à passer le fameux test de turing[10]. Bien que la communauté scientifique reste divisée quant à cette expérience, elle n'en reste pas moins une étape importante de l'intelligence artificielle, et plus particulièrement du Machine Learning[11].

## 1.2 Le vocabulaire propre au Machine Learning

Machine Learning (ML) ou apprentissage automatique : Il s'agit d'une branche d'étude de l'intelligence artificielle qui vise à permettre à une machine d'apprendre par elle-même à atteindre un certain but[12]. La différence fondamentale entre le Machine Learning et les modèles traditionnels est que le Machine Learning repose sur la recherche d'une corrélation et non d'un lien de causalité. Plutôt que d'établir les règles d'un modèle, l'algorithme trouve lui-même ces règles[13].On distingue plusieurs types d'apprentissages que nous détaillerons par la suite.

[6] https://www.quantmetry.com/single-post/2015/10/28/Une-petite-histoire-du-Machine-Learning
[7] https://fr.wikipedia.org/w/index.php?title=Histoire_de_l%27intelligence_artificielle&oldid=132868931
[8] https://www.quantmetry.com/single-post/2015/10/28/Une-petite-histoire-du-Machine-Learning
[9] http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/
[10] Le test de Turing, inventé par Alan Turing en 1960 consiste à dire que l'AI est atteinte lorsqu'un interlocuteur n'est plus capable de distinguer la conversation d'un humain d'une machine.
[11] http://www.reading.ac.uk
[12] https://fr.wikipedia.org/w/index.php?title=Apprentissage_automatique&oldid=134748222
[13] http://www.jeveuxetredatascientist.fr/quest-ce-que-le-machine-learning/




*Ces graphes représentent l'évolution des recherches sur Google pour les termes « Big Data » et « Machine Learning »[14]. Difficile donc de nier un certain engouement pour le sujet, et pourtant difficile de s'accorder sur le contenant du terme.*

Big data : Le Big Data est un terme récent, que l'on peut traduire par « données massives ». En 2001, le cabinet « Gartner » caractérise le Big Data par la règle des 3 « V » : volume, variété, vitesse[15]. Le volume représente la quantité de données traitée (plusieurs téraoctets, voir péta-octets[16]), un téraoctet représentant l'équivalent de 16 millions de photos (Gandomi and Haider, 2015) La variété fait référence aux types de données (texte, image, son, objets connectés, capteurs) et la vitesse fait référence à la fois à la génération et au traitement des données. Cette définition est parfois complétée par les « V » de « Veracity » et de « Variability » qui font respectivement référence à la qualité des données et à leur complexité.

Data-Mining : Le data-Mining, ou fouille de données, forage de données, cherche à extraire du savoir à partir de données. Le Machine Learning est donc une manière d'extraire des données parmi d'autres.

Deep-learning : le Deep-Learning fait référence à l'apprentissage automatique utilisant des algorithmes de réseaux de neurones avec un nombre important de couches. Il s'agit donc d'un

[14] https://trends.google.fr/trends/explore?date=all&q=big%20data
[15] https://fr.wikipedia.org/w/index.php?title=Big_data&oldid=137811293
[16] 1 péta-octet : 1024 TO

type particulier d'algorithme de  Machine Learning. Je reviendrai sur les réseaux de neurones par la suite.

Structure NoSql : Face aux volumes importants décrits ci-dessous, les systèmes classiques de bases de données en ligne présentent leurs limites. Les structures NoSql (Not-only Sql) permettent d'écrire et de lire les données plus rapidement, de stocker plus de données, et ce à un cout réduit. Le principe est de distribuer les données sur différents serveurs, contrairement aux bases de données classiques centralisées.

Data-visualisation : Si les capacités de calcul des ordinateurs ont permis l'exploitation de nombreuses variables, leur représentation devient délicate dès lors que l'on excède les 3 dimensions traditionnellement utilisées. Une branche des « data-sciences » s'est donc focalisée sur la représentation de ces données, essentielle à la fois dans le processus de « data-mining » et dans la communication des résultats. Les types de représentation dépendent des données à représenter (données temporelles, spatiales, multivariées, textuelles) et intègrent à la fois des techniques de distorsion des données (Keim, 2002).

Meta-learning : Le terme de Meta Learning est apparu en 1993 avec Chan et Stolfo. L'objectif de cette méthode est de structurer et d'optimiser le choix d'un algorithme en fonction d'un problème donné (Brazdil et al., 2008). Concrètement, il s'agit d'appliquer des algorithmes avec en variables d'entrées d'autres algorithmes afin de déterminer l'algorithme (et ses caractéristiques) optimal pour un problème donné.

## 1.3 Processus de l'analyse de données

L'analyse de données est un processus dont le succès n'est pas uniquement lié à la qualité des algorithmes utilisés. On distingue généralement 6 phases[17] :

- Définition du problème

L'analyse de données est un outil permettant de répondre à une problématique « métier ». Il est donc primordial d'énoncer clairement cette problématique. Cette phase permettra d'orienter le type d'algorithme adéquate (régression, classification, « clustering », détection d'anomalie).

- Collecte de données

[17] https://www.ibm.com/support/knowledgecenter/fr/SSEPGG_9.1.0/com.ibm.im.easy.doc/c_dm_process.html

Une fois la problématique traduite en termes d'analyse de données, il est nécessaire de collecter des données en quantité suffisante, mais aussi de qualité suffisante. Pour cela, les « data scientists » travaillent avec les spécialistes « métiers » pour explorer de nouvelles sources de données, puis utilisent les outils statistiques pour mesurer leur qualité. Durant cette phase, on utilisera aussi les outils de data-visualisation.

- Préparation des données

Cette phase consiste à « nettoyer » les données acquises, c'est-à-dire à traiter les données manquantes, reformuler et trier certains champs, normaliser les valeurs et globalement adapter les données afin qu'elles deviennent exploitables par l'algorithme.

- Modélisation et évaluation

Une fois le jeu de données établi, il est possible d'appliquer différents algorithmes pour une même problématique. Pour chacun d'eux, on évaluera alors sa performance au travers d'outils statistiques (que nous détaillerons par la suite). Il est possible d'agir sur les paramètres de l'algorithme pour affiner le résultat (par exemple profondeur d'un réseau de neurones).

- Déploiement

Un des enjeux de l'analyse de données est de pouvoir fournir des résultats dynamiques, et exploitables économiquement. Pour cela, il est nécessaire de déployer une solution en partenariat avec les services responsables des technologies de l'information (IT), mais aussi une interface (data visualisation). Il s'agit aujourd'hui d'un frein considérable au développement du Machine Learning dans les entreprises.

## 1.4 Principaux algorithmes utilisés

On distingue généralement deux types d'apprentissage :

### 1.4.1 Les modèles supervisés

Les classes (catégories) des exemples sont connues. On va donc alimenter l'algorithme avec des exemples tout en lui indiquant la catégorie (dans le cas d'un problème de classification) à laquelle l'exemple appartient. La phase de test consistera ensuite à lui présenter un élément et à mesurer son aptitude à « bien » le catégoriser. Les modèles supervisés probabilistiques apportent quant à eux une probabilité d'appartenance à une « classe ».

Parmi les modèles supervisés, on peut distinguer :

#### 1.4.1.1 Les modèles de classification

Comme évoqué précédemment, la finalité d'un tel modèle est de prédire si une donnée X appartient à telle ou telle catégorie définie au préalable. (par exemple « tendance baissière », « tendance haussière »). Dans cet exemple, et plus généralement lorsqu'il n'y a que deux choix possibles, on parle de classification binomiale. Les algorithmes les plus fréquemment utilisés sont :

- Two-class SVM (SVM) : Les algorithmes « Support Vector Machine » visent à établir une règle de séparation basée sur un hyperplan de marge optimale. On cherche donc à prédire Y dans {-1/1} à partir d'un vecteur de variables X1, X2, Xn[18]. Pour cela, l'algorithme optimise une fonction, qui prend en compte à la fois la qualité « prédictive » du modèle, mais aussi sa complexité. Cela permet d'éviter le phénomène de surentraînement (« overfitting») : un modèle qui s'est trop complexifié pour augmenter sa précision sur les données d'entrainement, mais qui s'avère moins efficace avec de nouvelles données (donc dans sa capacité à généraliser)[19]. Les SVM sont appréciés par leur solution « unique », interprétable comme un hyperplan.

- Two class decision trees (DT) : Les arbres de décisions représentent des algorithmes simples à comprendre. Dans un premier temps, on sépare les différents attributs des données. Puis on regarde si certains sont « purs », c'est-à-dire si leur seule connaissance permet d'estimer sans erreur la valeur à prédire. Dans le cas contraire, on va recommencer l'opération en séparant les données suivant un nouveau critère jusqu'à arriver à des situations « pures ». La question est donc de savoir dans quel « ordre » choisir ces critères. Pour cela, à chaque nœud de décision, on va mesure la « pureté » de la séparation suivant chaque critère restant. On utilise alors la notion d' « entropy » (la probabilité d'avoir d'une valeur « positive » serait inappropriée de par son aspect asymétrique).

$$Entropy = -P(+)\log_2 P(+) - P(-)\log_2 P(-)$$

[18] https://fr.wikipedia.org/w/index.php?title=Machine_%C3%A0_vecteurs_de_support&oldid=137517813
[19] http://www.math.univ-toulouse.fr/~gadat/Ens/M2SID/12-m2-SVM.pdf

Afin d'éviter le phénomène de surentrainement, on peut découper l'échantillon en sous-échantillons de manière aléatoire et étudier différents arbres de décisions. On parle alors de « Decision Forest » (DT).

- L'algorithme « perceptron » : c'est un réseau de neurones simplifié qui fait partie des classificateurs linéaires. Il fonctionne par itération sur chaque vecteur d'entrée (xi) et ajuste deux paramètres (un vecteur w et un biais b) si f(xi) différent de yi. A chaque nouvelle itération, on réalise le produit scalaire de w et xi auquel on ajoute b puis on compare le résultat à yi. Les itérations sont répétées jusqu'à trouver un vecteur w et un biais b qui permettent de prévoir correctement les valeurs yi.

- Réseaux de neurones (Two-class neural network- ANN) : Les réseaux de neurones permettent de faire face aux limites des modèles linéaires. Pour cela, ils vont utiliser un modèle de classificateur linéaire (comme le perceptron) tout en opérant une transformation entre deux neurones. Cela permet ainsi d'obtenir une classification plus complexe à partir de régressions logistiques simples. Ils sont particulièrement appréciés en finance par leur capacité à apprendre à identifier de manière autonome des relations non linéaires entre des données (Yoo, H. Kim, and Jan).

- Les Classificateurs Bayésiens Naïfs : ce sont des algorithmes qui reposent sur le théorème de Bayes. Il s'agit d'un classificateur linéaire basé sur l'hypothèse d'indépendance des variables d'entrées (d'où l'appellation « naïve », car ces variables peuvent cependant être corrélées). Un des avantages de ce genre d'algorithme est qu'il fonctionne avec peu de données d'entrées. En effet, si l'on considère les variables comme indépendantes, les matrices de covariances deviennent inutiles.

#### 1.4.1.2 Les modèles de régression

Il s'agit des modèles les plus utilisés. C'est pourquoi nous ne détaillerons pas les spécificités de chaque modèle. On retiendra qu'un modèle de régression se distingue d'un modèle de classification par la variable de sortie qui peut prendre des valeurs numériques continues. Parmi les modèles les plus connus, on retrouve les régressions linéaires (très utilisées en économétrie). Par ailleurs, la plupart des algorithmes de classification peuvent être utilisés pour des régressions, on parle alors de régression linéaire bayésienne, de modèle neuronal pour la régression, d'arbres de décisions pour régression.

### 1.4.2 Modèles non supervisés

A contrario, un modèle non supervisé sera quant à lui alimenté uniquement par des exemples, et créera lui-même les classes qui lui semblent les plus judicieuses (clustering) ou des règles d'associations (algorithmes Apriori). L'algorithme K-moyennes (K-means) permet de comprendre facilement le concept de classification non supervisée.

Cet algorithme est un algorithme de « clustering », c'est-à-dire qu'il va créer des catégories à partir des observations. Pour cela on détermine un entier k correspondant au nombre de partitions (groupes) souhaitées. On va ensuite chercher à minimiser la distance entre chaque point et le point représentant la moyenne d'une partition donnée.

L'algorithme se déroule comme suit :

- On choisit au hasard des points qui représentent la position moyenne des k partitions.

Puis on répète les opérations suivantes jusqu'à obtenir une convergence :

- On attribue à chaque observation sa partition la plus proche (en comparant sa distance à chaque point précédemment positionné).
- Une fois les observations regroupées par partitions, on remet à jour la position moyenne[20].

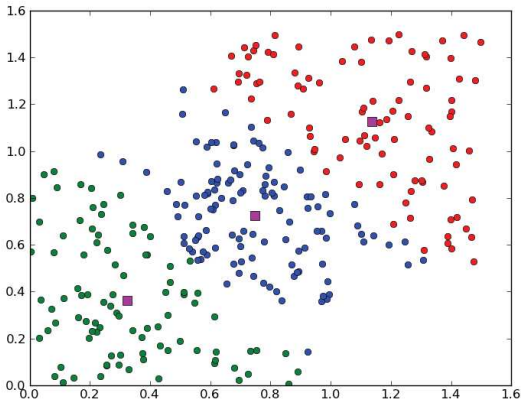



FIGURE 1EXEMPLE DE CLUSTERING[21]

[20] https://fr.wikipedia.org/w/index.php?title=K-moyennes&oldid=131998110
[21] https://dzone.com/articles/k-means-clustering-scipy

### 1.4.3 Fléau de la dimension et méthodes de réduction des variables

Le fléau de la dimension fait référence aux problématiques rencontrées lorsque l'on utilise des espaces avec de nombreuses dimensions (dans notre cas, des données avec de nombreux attributs)[22]. Cela se traduit, soit par une perte de précision (la représentation d'un même nombre de données est plus éparse si l'on augmente les dimensions) soit par des quantités de données nécessaires plus importantes. Pour y remédier, différentes techniques de réductions de variables sont possibles dont l'analyse en composantes principales.

## 1.5 Evaluation des modèles et précision des résultats

Dans cette partie, nous rappellerons à la fois les mesures utilisées pour les algorithmes de classification et les régressions linéaires.

### 1.5.1 Sensibilité, spécificité et valeur prédictive positive

Prenons l'exemple de la prévision d'une « hausse » ou d'une « baisse » d'un indice. Une fois l'algorithme entrainé, nous allons le tester sur une base de donnée spécifique (ce qui génèrera les résultats F(X)), dont les labels « hausse » ou « baisse » sont connus (Y). Cela permettra de déterminer :

| | Y = « hausse » (réalité) | Y = « baisse » (réalité) | Total |
|---|---|---|---|
| F(X) = « hausse »<br>(estimation de l'algorithme) | 12<br>(vrais positifs) | 5<br>(faux positifs) | **17** |
| F(X) = « baisse »<br>(estimation de l'algorithme) | 7<br>(faux négatifs) | 16<br>(vrais négatifs) | **23** |
| **Total** | **19** | **21** | |

On définit alors la Sensibilité (ou rappel / « recall ») comme le rapport de vrais positifs parmi l'ensemble des positifs à détecter (12/19) et la spécificité comme la proportion des vrais négatifs parmi l'ensemble des négatifs à détecter (16/21). On parle aussi de **précision**, comme le nombre de valeurs correctement estimées « positives » sur le nombre de valeurs estimées positives (12/17). Cet indicateur reviendra souvent dans les articles étudiés de la partie 4.

[22] https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales

Afin de combiner ces deux critères, on utilise le F1-score qui n'est autre que la moyenne harmonique entre la précision et le rappel : plus l'algorithme est performant, et plus le F1-score se rapprochera de 1.

$$F_1 = 2 \times \frac{1}{\frac{1}{rappel} + \frac{1}{précision}} = 2 \times \frac{précision * rappel}{précision + rappel}$$

### 1.5.2 Courbe ROC

La « courbe ROC » (« Receiver Operationg Characteristic), ou fonction d'efficacité du récepteur » permet de mesurer la performance des prédictions d'un algorithme de classification[23].On représente en abscisse le taux de faux positifs (le modèle les prévoit comme «positif », à tort) et en ordonnée les vrais positifs. On peut retenir que les courbes situées au dessus de la diagonale permettent d'améliorer la prédiction par rapport à un choix au hasard.

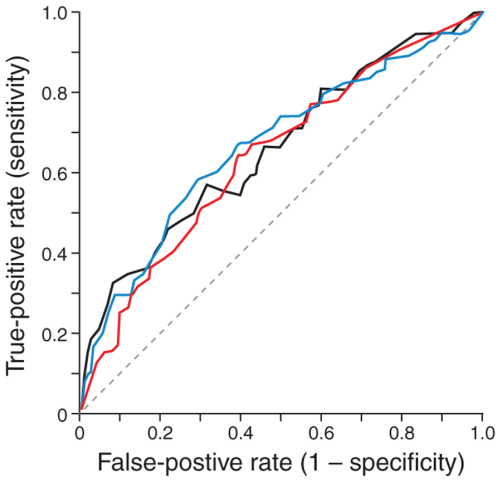



FIGURE 2 EXEMPLE DE COURBE ROC[24]

Afin de pouvoir quantifier cette mesure, on utilise généralement la mesure de l'aire sous cette courbe. Evidemment plus celle-ci est importante, plus l'algorithme est performant.

### 1.5.3 Outils statistiques pour les régressions linéaires

Bien que ce travail de recherche se focalise sur les algorithmes plus « récents » de Machine Learning, il me semble important de rappeler les principaux tests statistiques qui permettent de mesurer la performance des régressions linéaires. Cela permettra entre autre de mieux

[23] https://fr.wikipedia.org/w/index.php?title=Courbe_ROC&oldid=126274338
[24] http://www.nature.com/nbt/journal/v28/n5/fig_tab/nbt0510-444_F1.html

comprendre les résultats de certaines études reposant en partie sur des régressions linéaires multiples.

- P-value : T-stat et p-value : ils permettent de mesurer la significativité d'un coefficient. On peut retenir qu'un coefficient est significatif avec un intervalle de confiance de 95% lorsque le t-stat est supérieur à 1,96 ou lorsque la p-value est inférieure à 0,05.
- Coefficient de détermination : Permet de mesurer la qualité du modèle. Il s'agit « du rapport entre la variance expliquée et la variance totale ». Le R2 est donc compris entre 0 et 1. Il existe aussi un coefficient de détermination « ajusté » qui prend en compte le nombre de variables du modèle.

# 2 Le Machine Learning en finance

## 2.1 De nouvelles sources de données

On estime que les quantités de données disponibles dans le monde doublent tous les 1,2 ans (Manyika et al., 2012) et que cela justifie en grand partie le regain d'intérêt du Machine Learning. L'usage de données en finance n'est pourtant pas un phénomène récent : les méthodes d'évaluation de risque, de valorisation, de pricing s'appuient sur des données, principalement chiffrées.

Un des changements majeurs de ces dernières années concerne la nature nouvelle des données. On peut entre autre citer :

- Le texte : nous étudierons par la suite des études portant sur les publications de sites d'actualités et des réseaux sociaux. L'objectif étant généralement d'extraire un sentiment de ces données, nous détaillerons par la suite les différentes approches possibles.
- L'image: largement utilisée en météorologie, mais aussi en agriculture (marché des matières premières), ou encore pour mesurer en direct la fréquentation de centres commerciaux[25]. Il est aujourd'hui possible d'extraire des connaissances à partir de photos ou de vidéos (e.g. reconnaissance de visages ou de marques de produits sur Facebook).
- Conversations : des algorithmes spécifiques existent aujourd'hui pour analyser des conversations (à la fois écrites et orales). Cela se traduit par des chatbots[26] : des « robots » capables d'interagir avec un humain au travers d'une discussion.
- La géo localisation : L'usage de la géo localisation permet de mesurer l'affluence dans un lieu, et plus généralement les habitudes d'un utilisateur (lieu de travail, résidence, type de trajet etc…)
- Navigation internet : il est aujourd'hui possible de suivre le mouvement de la souris lors d'une navigation, afin de juger de l'attitude d'un prospect sur un site e-commerce, ou encore du temps passé sur chaque slide d'une présentation.

[25] https://www.nytimes.com/2017/03/28/business/dealbook/blackrock-actively-managed-funds-computer-models.html
[26] e.g. Siri IOS / Cortina sur Microsoft

- Les données d'objets connectés : C'est particulièrement valable dans le domaine des assurances. Boyer (2015) fait entre autres référence aux boitiers connectés pour les automobiles (Google Car, Tesla) mais aussi aux capteurs biométriques pour la santé (Health Apple). Ce phénomène de collecte de données à la fois « biologiques, physiques, comportementales ou environnementales » est aussi appelé « soi quantifié » (« Quantified Self », Swan , 2013).

## 2.2 Applications en gestion de risque et assurance

### 2.2.1 Tarification des assurances

Le Big Data permet une tarification plus précise, et ce, à différents niveaux. On peut tout d'abord faire référence à la précision du « zoning »[27], aux pratiques de l'assuré (nombre de kms réellement parcourus ou type de conduite dans le cas d'une assurance automobile). Boyer (2015) évoque par exemple l'étude des corrélations entre les mouvements d'un compte courant et les accidents, par exemple les montants payés en CB Cependant, l'usage des données pour une segmentation précise des clients « à risques » et la tarification spécifique qui peut en découler mène évidement à un débat éthique et réglementaire sur ces pratiques.

Un autre champ du Machine Learning exploité en assurance concerne les détections de fraudes : l'Electronic Fraud Detection( EFD). Considérant qu'entre 125 et 175 milliards de dollars sont perdus chaque année sur le système de santé américain, Travaille et al. (2011) montrent que des algorithmes supervisés permettent de mieux prévenir les fraudes.[28] Des travaux similaires ont été réalisés sur les assurances automobiles, estimant entre 10 et 20% les réclamations frauduleuses[29]. Concernant ce dernier secteur, Viaene et al. (2002) montrent que des algorithmes relativement simples (comme les régressions linéaires logistiques ou les réseaux bayésiens naïfs) mais robustes et nécessitant peu de données d'entrainement sont plus performants que des algorithmes plus complexes (comme l'algorithme C4.5 issu des arbres de décisions).

[27] Tarification suivant les zones géographiques
[28] Pour plus d'informations sur le sujet, voir H.Joudaki & al, « *Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature* » 2014
[29] Voir M. Aris & al. « *Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims* », 2002

### 2.2.2 Prévision des défaillances d'entreprises

Ben Jabeur et Fahmi (2013) étudient la prévision des défaillances sur un échantillon de 800 entreprises à partir de 33 ratios. Ils comparent la régression des moindres carrés partiels (régression PLS) [30] avec une approche SVM et aboutissent à des résultats similaires (96,5% de bon classement pour l'approche PLS contre 94,9% pour la méthode SVM un an avant la défaillance, supériorité de l'approche PLS 3 ans avant la défaillance, supériorité de la méthode SVM 2 ans avant la défaillance). Y. Alici[31] montre quant à lui une supériorité des réseaux de neurones par rapport aux méthodes d'analyse discriminante et de régression logistique pour la prédiction des faillites.

### 2.2.3 Evaluation du risque de crédit

L'évaluation du risque de crédit est une problématique à la fois des banques de détail (crédits aux particuliers et entreprises) et de finance de marché (e.g. Credit Default Swaps, gestion de collatéral). On peut séparer deux types d'évaluation : a) l'évaluation lors d'une demande de crédit b) l'évaluation de risque de défaut durant la vie du crédit (Khashman, 2010). Les études dans ce domaine s'appuient sur des approches statistiques classiques (régressions logistiques), des arbres de décisions et les réseaux de neurones, qui semblent là aussi les plus utilisés et les plus performants (Baesens et al., 2003). Cependant, la nécessité d'expliquer la raison d'un accord ou surtout d'un refus de crédit a orienté la recherche vers l'extraction de « règles » à partir de réseaux de neurones entrainés (algorithmes dits de « décompositions » : Neurorule, Trepan, Nefclass). A titre d'exemple, Baesens et al. (2003) utilisent différents réseaux de neurones sur un jeu de 1000 « cas » de crédits labélisés « accept » ou « reject » (entrainement supervisé). Chaque cas est constitué de 20 variables (numériques ou catégoriques e.g. type de travail, âge, nombre de crédits en cours). Ils obtiennent un modèle prédictif avec 83,6% de précision. Cette étude souligne d'une part la nécessité d'avoir des données suffisantes pour l'entrainement et l'évaluation du modèle mais aussi la difficulté pour normaliser des variables d'entrées.

### 2.2.4 Détection de fraudes financières

Dans leur article, A. Abbasi et al. proposent une méthode de détection des fraudes basée sur l'étude d'informations publiques en utilisant le meta learning (Abbasi et al., 2012). Pour cela,

[30] La régression PLS est issu de la régression et de l'analyse en composantes principales.
[31] Y. Alici « *Neural networks in corporate failure prediction: The UK experience* », 1996

ils s'appuient sur les publications trimestrielles et annuelles des entreprises en étudiant 12 ratios financiers. Le modèle proposé s'avère plus performant que les modèles développés jusqu'alors avec un AUC entre 0,864 et 0,922 suivant les ratios utilisés. Ngai et al. (2011) synthétisent les algorithmes suivant le type de fraudes à détecter (fraudes de cartes de retraits, fraudes d'assurances, fraudes d'entreprises). Concernant la détection de fraudes d'entreprises, les réseaux de neurones et les régressions logistiques semblent être les plus utilisés (Yue et al., 2007). La principale limite des réseaux de neurones dans ce domaine reste l'effet « boite noire » et donc l'impossibilité de comprendre l'algorithme (i.e. les facteurs de fraudes). En terme de performance, Wang (2010) synthétise les résultats de différentes études sur le sujet. Les taux de précision atteignent les 90% (84% pour Yue et al. (2007)), mais l'auteur souligne la difficulté d'une telle mesure sachant que les « labels » des séries d'entrainement dépendent des analyses traditionnelles passées (des entreprises non détectées auparavant comme frauduleuses peuvent cependant l'avoir été…). Si les ratios financiers sont les plus utilisés (e.g. inventaires / ventes, dettes / actif, résultat net / actif), Yue et al. (2007) proposent d'intégrer des données liées à la gouvernance ainsi qu'aux audits réalisés.

Les études de fraudes[32] concernent aussi les marchés financiers. On peut distinguer différents types de fraudes dont les fraudes de rendement (e.g. chaines de ponzi, frais de courtages injustifiés), les fraudes de brokers (deals non autorisés à partir des titres de clients, deals réalisés en dehors des heures de marché) ainsi que les manipulations de marché (Golmohammadi et Zaiane, 2012) Les données utilisées sont à la fois variées (articles, messageries de traders, bases de données répertoriant les opérations) et volumineuses (plus de 5000 transactions par seconde sur le NASDAQ, Golmohammadi and Zaiane (2012)).

## 2.3 Machine Learning et gestion de portefeuille

### 2.3.1 Développement du trading algorithmique

BlackRock est aujourd'hui la société de gestion la plus importante au monde, avec plus de 5000 Md d'euros en gestion fin 2016[33]. Face au développement des Exchante Traded Funds (ETF) ou trackers, réputés pour leurs faibles coûts, BlackRock a investi dans le trading algorithmique. En mars 2017, le fond annonce viser 30 Md€ investis dans des fonds algorithmiques, une décision qui se traduit aussi par le licenciement d'une trentaine de « stock

[32] K. Golmohammadi & al (2012) recensent 205 études sur le sujet entre 2001 et 2012, dont la majorité à partir de 2008.
[33] https://fr.wikipedia.org/w/index.php?title=BlackRock&oldid=136776518

pickers ». Les « nouveaux » gestionnaires d'actifs *« might buy (or sell) Walmart's stock on the basis of a satellite feed that reveals how many cars are in its parking lots as opposed to an insight gleaned from the innards of the retailer's balance sheet. »*[34]. Si l'amalgame entre analyse de données et trading haute fréquence est fréquent dans la presse, il convient cependant de distinguer ces deux approches. Le trading haute fréquence peut en effet s'appuyer sur de l'analyse prédictive (souvent conjointement des algorithmes issus de la théorie des jeux), mais l'analyse prédictive peut aussi être un outil d'aide à la décision pour la gestion de portefeuille classique. Enfin, le trading haute fréquence s'appuie sur des volumes importants et des opérations très fréquentes, jouant plus un rôle de « market maker » que de gestionnaires de fonds.

### 2.3.2 Quel avenir pour les gestionnaires de fonds ?

L'engouement pour le Big Data ne se traduit pas tout le temps par des succès au niveau des fonds d'investissement. En effet, l'analyse de données reste un procédé complexe et coûteux, en développement et dont les résultats sont encore mitigés (Nardo et al., 2016). Plusieurs fonds semblent d'ailleurs en faire les frais : le fond Big Data de Catana Capital a perdu 3,9 % de sa valeur durant ses 3 premiers mois (juillet 2016), contre 0,8 % seulement pour le DAX. Chez BlackRock aussi, les fonds quantitatifs n'ont pas réussi à atteindre leurs objectifs en 2016[35].

En dehors de ces performances contestées, les performances de l'analyse prédictive dépendent encore beaucoup de l'expertise des analystes. C'est à la fois vrai dans le choix des variables d'entrées (analyse fondamentale ou technique) mais aussi dans le traitement de ces dernières (lexiques spécifiques pour le milieu financier (Renault, 2017). Comme dans de nombreuses industries, l'analyse prédictive s'annonce plus comme une aide à la décision qu'un remplacement des acteurs.

[34] https://www.nytimes.com/2017/03/28/business/dealbook/blackrock-actively-managed-funds-computer-models.html

[35] https://www.lesechos.fr/finance-marches/gestion-actifs/0211720097114-comment-le-big-data-simpose-dans-la-gestion-dactifs-2059492.php#Xtor=AD-6000

# 3 Le machine Learning appliqué à la prévision des cours

Dans cette partie, nous étudierons l'impact des nouveaux algorithmes de Machine Learning sur la question de la prédictibilité des cours. La littérature sur ce sujet est abondante. J'ai pour ma part étudié 41 études citées dans 7 synthèses (Mclean and Pontiff (2016), Nardo, Petracco-Giudici, and Naltsidis (2016), Kearney and Liu (2014), Yoo, Kim, and Jan (2005), Baker and Wurgler (2007), Lawrence (1997), Tsai and Wang, S-P.). publiées entre 1988 et 2017 afin de mettre en avant les variables aléatoires étudiées, les sources de données, les algorithmes utilisés et les performances réalisées. C.-F. Tsai and Wang, S-P. synthétisent les résultats de 10 études ciblées sur les marchés asiatiques utilisant des algorithmes de Machine Learning. Pour ma part, j'ai inclus certaines études aux approches « classiques » de régression multi variables afin de pouvoir comparer les résultats (liste complète en Annexe B).

## 3.1 Hypothèse d'Efficience des Marchés et stratégies

### 3.1.1 Une remise en cause de l'HEM

Selon Fama (1970), un marché est efficient si les prix reflètent toute les informations disponibles. Il distingue 3 formes d'efficience qui dépendent de la nature de l'information disponible :

- L'efficience faible : seuls les prix passés d'un actif sont pris en compte dans son prix à un instant t.
- L'efficience semi-forte : les prix passés ainsi que les informations publiques sont prises en compte dans le prix à un instant t.
- L'efficience forte : toutes les informations, à la fois publiques et privées sont prises en compte dans le prix à un instant t.

L'efficience forte se traduit par l'absence de gain anormal (ajusté du risque), un sujet fréquemment étudié au travers des études d'événements. La finance comportementale considère quant à elle que les prix peuvent être influencés par les sentiments des investisseurs.

Les algorithmes de Machine Learning alimentés par des données publiques remettent ainsi en cause (pour un temps du moins) l'efficience semi-forte des marchés. En effet, si un algorithme arrive à « prévoir » avec une précision supérieure à 50% l'évolution d'un prix, c'est qu'une part de l'information disponible n'a pas été exploitée. Evidemment si cette « anomalie » devient « connue » par un nombre suffisant d'acteurs, les prix reviendront à l'équilibre (aux couts d'arbitrage près). Grossman et Stiglitz (1980) considèrent cependant que des légères surperformances sont possibles et qu'elles « payent » le temps passé par les investisseurs dans leurs recherches d'informations.

## 3.1.2 Evolution des approches de trading

### 3.1.2.1 L'analyse fondamentale

L'analyse fondamentale considère que la valeur d'un actif est égale à l'actualisation de ses flux futurs à un taux qui dépend du risque propre à ce titre. Cette analyse repose essentiellement sur les données comptables publiées par la société, son prévisionnel pour les années à venir et sur des facteurs économiques impactant la société. Différentes méthodes permettent de réaliser ce genre d'analyse parmi lesquelles les Discounted Cash-Flows, Gordon- Shapiro, ou encore les ratios financiers. De manière plus générale, l'analyse fondamentale fait appel à la macro-économie et à la stratégie d'entreprise. Les investissements réalisés sont généralement de moyen-long terme.

### 3.1.2.2 L'analyse technique ou chartiste

L'analyse technique utilise comme données uniquement les cours historiques d'un titre. Cette analyse repose sur le fait que des schémas se répètent et peuvent donc anticiper les mouvements d'un titre. De nombreuses « figures » sont alors reconnaissables, et les outils tels que les moyennes mobiles ou les bougies japonaises aident à la décision.

L'effet Momentum se caractérise par une plus grande probabilité des titres ayant eu des performances positives (négatives) par le passé à avoir des performances positives (négatives) dans le futur. Les stratégies associées consistent donc à ré arbitrer régulièrement suivant les performances des derniers mois[36].

L'analyse technique est globalement remise en cause par les études académiques, mais suscite toujours un intérêt, en particulier chez les tradeurs novices. Certaines études semblent

[36] http://www.bsi-economics.org/204-biais-cognitifs-effet-momentum

expliquer certains « succès » de ces stratégies par un phénomène auto-réalisateur du à un nombre important d'acteurs adoptant des stratégies similaires (Menkhoff, 1997).

## 3.2 Variables étudiées

Comme le montre le tableau ci-dessous, la plupart des articles étudient la prédictibilité d'indices. On retrouve en première place les indices américains (S&P500, Dow Jones, NASDAQ) mais aussi les indices asiatiques (KOSPI, Taiwan Stock Index, TOPIX)[37]. Il s'agit donc de valeurs particulières puisqu'ils représentent un panier d'actions, mais ils ont l'avantage de jouir d'une grande liquidité ainsi que d'une représentativité globale de l'économie d'un pays (cas des études de tendance sur les sentiments). Concernant les études sur des titres individuels, les auteurs étudient en général les cours de plusieurs titres (plus d'une centaine en moyenne). On peut aussi distinguer les approches sectorielles (agriculture pour Ng and Khor (2017), industrie du ciment pour Fallahi, Shaverdi, and Bashiri (2014), électronique pour Changa and Wangb (2013), Internet pour Tumarkin and Whitelaw (2001)). Enfin seule une étude (une des premières utilisant les réseaux de neurones) se concentre sur un seul titre (IBM, (White, 1988)).

| | |
|---|---|
| Autres | 2 |
| Indice | 25 |
| Titres individuels | 13 |
| **Total général** | **40** |

*Répartition des variables étudiées*

Il est intéressant de noter que les zones couvertes sont relativement restreintes et concentrées sur les Etats-Unis et certains pays asiatiques (Corée du sud et Taiwan principalement). Aucune des études que j'ai étudiées ne porte sur les marchés européens par exemple.

| | |
|---|---|
| Amérique du nord | 23 |
| Asie | 16 |
| **Total général** | **39** |

*Répartition des marchés étudiés*

Enfin, concernant les durées d'observations, la durée moyenne des études est de 7,5 ans (elle varie d'un mois (Schumaker and Chen) à 26 ans (Gensai, 1996). La plupart des rentabilités

[37] Voir annexe pour le détail par étude

observées sont journalières, tout comme l'horizon des prévisions. Certaines approches fondamentales s'appuient sur des données trimestrielles. On retiendra que les variables étudiées sont assez « normées », probablement par soucis de comparabilité.

## 3.3 Stratégies et sources de données

Je me suis ensuite intéressé aux stratégies sous-jacentes aux études. Pour cela j'ai classifié les études suivant les données utilisées entre stratégies « techniques », « comportementales » et « fondamentales », ainsi que des stratégies hybrides.

| | |
|---|---|
| Approche technique | 13 |
| Comportementale | 10 |
| Comportementale + fondamentale | 3 |
| Comportementale + technique | 3 |
| Fondamentale | 11 |
| **Total général** | **40** |

*Répartition des stratégies*

### 3.3.1 Approches techniques

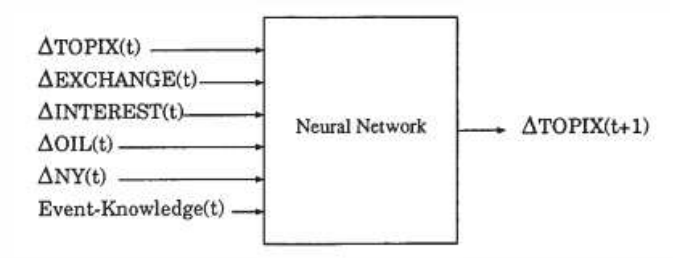
Un premier type d'étude, qui vient remettre en cause la forme faible d'efficience des marchés s'appuie sur l'étude des cours passés (en général sur les rentabilités quotidiennes, voir intra-day). Ce sont souvent des données étudiées par des chercheurs issus de formations « informatiques », souhaitant se concentrer sur l'étude et la comparaison des performances des algorithmes[38]. Les données utilisées sont les cours historiques, mais aussi des indices techniques (e.g. moyennes mobiles, Relative Strengh index, mesures des effets momentum).

### 3.3.2 Approches fondamentales

D'autres études étendent leurs données à des ratios financiers et économiques principalement issus des services d'information financière (Reuters, Bloomberg, Datastream). Parmi les ratios fréquemment utilisés on retrouve le rendement, le « Return on equity », le « Price to book ratio », le taux de marge ainsi que le ratio de liquidité. On peut citer l'approche originale de Changa and Wangb (2013) qui incluent d'une part la valorisation par le modèle d'Ohlson en variable d'entrée mais aussi des variables liées à la gouvernance (CEO shareholding ratio,

[38] Voir par exemple les 3 études de K.Kim basées sur les mêmes données, mais avec des algorithmes différents.Kim (2004),Kim (2003),Kim and Han (2000)

composition de la direction). D'autres études intègrent aussi des variables économiques (taux d'intérêts, indices de consommation pour Chen, Leung, and Daouk (2003),prix des matières premières, taux de changes pour Enke and Thawornwong (2005)). C'est le cas de Kohara qui s'appuie à la fois sur des événements passés (extraits et traités manuellement à partir de journaux) et différents indices tels que les taux de change, les intérêts actuels, le prix du baril de pétrole ou encore le Dow-Jones pour prévoir l'évolution du TOPIX (TOkyo stock Price IndeX). Il montre que l'intégration des évènements permet d'améliorer de 40% la rentabilité d'une stratégie basée sur son algorithme.




*Variables d'entrées utilisees par Kohara et al*

Une troisième famille d'étude, entre les stratégies fondamentales et comportementales analysent les articles traditionnels issus de la presse spécialisée, par exemple les articles du Wall Street Journal (Tetlock, 2007) ou du Financial Times[39], ce qui permet d'assurer une certaine qualité et homogénéité des textes (Nardo et al., 2016). De plus en plus de sites internet proposent directement des flux d'articles issus de sources variées: PRNewsWire[40], Yahoo Finance[41].

### 3.3.3 Approches comportementales

Une première série d'études basées sur la mesure de « sentiment » s'appuie sur des indices publiés propres aux marchés financiers (MArketPsych Indices TRMI pour Sun, Najand, and Shen, (2016)) ou issus des réseaux sociaux (Facebook's Gross National Happiness pour Karabulut (2013)). Il s'agit donc encore de données « quantitatives » intégrées dans les modèles, qui plus est reflétant un sentiment « général ».

[39] Alanyali & al, 2013
[40] http://www.prnewswire.com/
[41] https://finance.yahoo.com/news/provider-businesswire/

Une deuxième série d'études s'appuient uniquement sur le nombre de visites de pages internet. On retiendra l'étude de Moat et al. (2013) qui mesurent les pages vues ou éditées sur Wikipédia relatives à certains titres ou au secteur financier en général.

Mais les données les plus utilisées ces dernières années proviennent des réseaux sociaux. On peut distinguer les réseaux sociaux spécialisés dans le trading comme StockTwits (Renault, 2017) Ragging Bull (Tumarkin and Whitelaw, 2001), (Antweiler and Frank, 2004), Yahoo Finance (Kim and Kim, 2014), CAPSAvery (Chevalier, and Zeckhauser, 2016), HotCopper (Leung and Ton, 2015) . A titre d'exemple, T. Renault étudie les corrélations entre les messages et l'évolution d'un indice à partir de 60 millions de messages pour une période de 5 ans. On peut noter l'importance des « labels » explicites sur ces forums (« haussiers », « baissiers ») qui permettent de facilement classifier les messages. Concernant les réseaux non spécialisés, Twitter est le plus utilisé, en particulier grâce à la standardisation des messages publiés et à ses volumes importants[42]. Contrairement aux articles traditionnels, il est plus difficile d'analyser ce type de message, d'une part car le vocabulaire utilisé est parfois spécifique ou abrégé, et d'autre part car le ton des phrases (ironie) est plus difficilement interprétable car moins standardisé.

On retiendra aussi que la plupart des études comportementales ou fondamentales utilisent aussi les cours passés afin d'augmenter la précision des modèles. Cependant, comme nous l'avons vu précédemment, l'augmentation de variables d'entrées nécessite des jeux de données très importants. Pour faire face à ce problème, différentes méthodes permettent de réduire les variables d'entrées dont l'analyse en composantes principales (Datta Chaudhuri, Ghosh, and Eram, 2015).

## 3.4 Algorithmes utilisés

### 3.4.1 Analyser le texte

Les premières analyses de texte sur le web se sont concentrées sur le nombre d'apparition d'un mot ou de caractères clefs (en particulier le nom des symboles)(Tetlock, 2007). A ce niveau là, seule une corrélation entre le nombre d'apparition de mots clefs et le volume d'échange du titre en question est observable (Nardo et al., 2016).

[42] H. Mao & al.Bollen, Mao, and Zeng (2011) ont collecté 10 millions de tweets publiés par 2,7 M d'utilisateurs sur une période de 10 mois.

On peut distinguer deux grandes méthodes pour la mesure de l'humeur issue d'un texte : l'usage d'un dictionnaire et une approche par les algorithmes de Machine Learning. Parmi les dictionnaires utilisés, on retrouve le dictionnaire Harvard IV (Tetlock, 2007), le Loughran and Macdonald (Engelberg, Reed, and Ringgenberg, 2012), plus utilisés pour l'analyse de médias classiques.

Dans son étude, T.Renault construit deux lexiques spécifiques : le premier est un lexique dont les mots sont pondérés à partir de 750 000 messages étiquetés « haussiers » ou « baissiers le deuxième est une classification manuelle sur les termes revenant plus de 75 fois dans les messages. ». Il les compare ensuite aux dictionnaires Harvard IV ainsi qu'au dictionnaire de Loughran et Macdonald. Il ressort de cette étude que seul un lexique spécifique ou un algorithme de classification (Maximum Entropy Classifier) permettent d'obtenir une estimation correcte de l'humeur des messages. (76,36% avec un lexique spécifique, 75,16% pour un algorithme de classification contre 58,29% pour un dictionnaire Harvard IV.) Cela confirme les résultats de Loughran et McDonald qui montrent que 74% des classifications « négatives » générées par les dictionnaires Harvard ne sont pas valables dans le domaine de la finance.

(Nardo et al., 20016) listent d'autres outils de classifications utilisés pour déterminer l'humeur d'un texte, tel que « Opinion Finder », Google Profile of Mood States » , WorldTracker (qui permet d'identifier les mots les plus cherchés en rapport avec un mot clef), Senti WN (donne une mesure positive / négative) mais aussi les « boosted decision trees » ou encore les classifieurs bayésiens naifs.

Schumaker and Chen développent un outil spécifique à partir d'un logiciel (Arizona Text Extractor). La manière la plus simple de représenter un article est de générer un vecteur où les mots sont indexés puis pondérés. Il y a donc une première phase de sélection des mots. Pour cela, différentes méthodes sont possibles :

- Sac de mots « bag of words » : on enlève simplement des mots de liaison définis auparavant dans une liste.
- « noun phrases » : Cette méthode est une extension du « bag of word » qui permet d'identifier le type de mot (noms propres, noms communs, verbes, adjectifs, grâce à la reconnaissance de mots « types » de liaison.

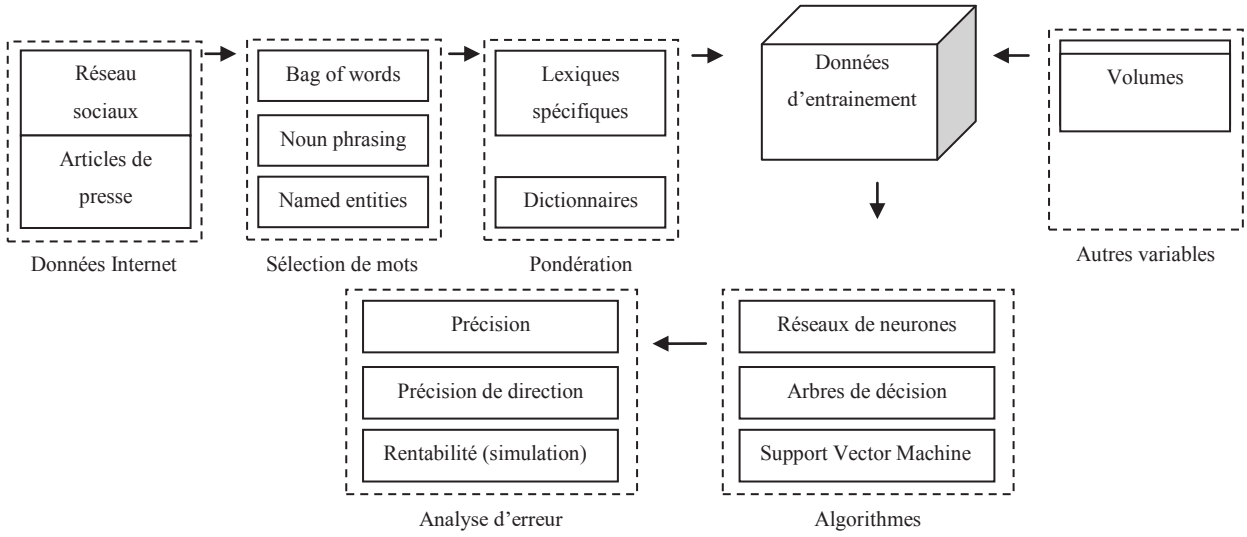
- « named entities » : Il s'agit cette fois ci de catégoriser des structures de phrases suivant des catégories établies au préalable ( par exemple des dates, des lieux, des personnes etc…) Schumaker and Chen

La conclusion des auteurs est que le modèle « named entities » s'avère trop restrictif dans ses choix alors que le « noun phrases » contient trop de bruits. Les auteurs développent une solution hybride qui ajoute à la sélection « named entities » les noms propres issus de la méthode « noun phrases ». Ils obtiennent ainsi les meilleures performances à la fois en termes de direction et de rentabilité. Ce qui encore une fois confirme la nécessité d'adapter et de spécifier les dictionnaires utilisés en fonction du secteur étudié. En terme de précision, les classifications binaires obtiennent de bons résultats, (76,6 % pour Renault (2017), 70,3% pour Koppel and Shtrimberg (2006)).

L'analyse du texte et surtout de la communication est un champ de recherche très dynamique. Parmi les pistes de développement, l'usage de cartes cognitives[43] associées à des réseaux de neurones semble prometteuses (Hong and Han, 2004).

### 3.4.2 Choix des algorithmes

Une fois le texte analysé, il est généralement combiné à d'autres données, en particulier avec des cours historiques afin de déterminer ses corrélations avec les rentabilités futures. Dans cette phase, les choix d'algorithmes sont divers.




*Process de l'analyse de données appliquée aux articles et textes*

[43] Réseaux entre des concepts et de liens de causalité qui visent à modéliser le comportement humain.

Les algorithmes de réseaux de neurones sont les plus utilisés. Leur première utilisation dans ce domaine remonte à 1988 avec une étude sur la prédictibilité des cours d'IBM[44]. Les réseaux de neurones sont devenus plus performants que les méthodes statistiques (comme l'analyse factorielle discriminante) dans les problèmes de classification. C'est le cas de l'étude de Lawrence[45] qui obtient une précision de 92% dans ses prévisions avec un réseau de neurones contre 60% avec des régressions, ou encore de Yon & al[46]. qui obtient une précision de 91% avec des réseaux de neurones contre 74% en utilisant l'analyse factorielle discriminante (Yoo, H. Kim, and Jan).

De nombreuses études, en particulier issues des chercheurs en informatique ont visé à comparer les algorithmes entre eux. Kim (2003) montre une meilleure performance avec les modèles SVM (en obtenant 57,83 % de bonnes prédictions contre 54,7 % avec un réseau de neurones « back propagation » et 51,9% avec un arbre de décision) Il est cependant important de distinguer le type de variables d'entrées pour évaluer différents algorithmes. Kim utilise uniquement des indicateurs techniques (12 au total) relatifs aux prix passés.

Globalement, on observe une tendance à utiliser plusieurs algorithmes. C'est le cas par exemple des algorithmes génétiques qui sont souvent utilisés en amont des réseaux de neurones pour sélectionner les variables d'entrées et les caractéristiques du modèle (nombre de couches).Yoo, H. Kim, and Jan propose en 2004 un modèle hybride utilisant les algorithmes génétiques pour sélectionner les variables d'entrées d'un arbre de décision, qu'il compare à des arbres de décisions classiques. A partir des mêmes indices techniques qu'il avait utilisé en 2003, Kim montre une meilleure performance de son modèle hybride par rapport aux arbres de décisions classiques et aux SVM (obtenant alors 60,7 % de bonne prédictions contre 52,14% avec un arbre de décision classique) . Kim (2004), Tsai and Wang appliquent eux aussi des combinaisons (ANN+DT, DT+DT) après avoir réalisé une sélection des variables grâce à une analyse en composantes principales. Voilà les résultats qu'ils en tirent :

| Type d'algorithme | Précision observée |
|---|---|
| Arbres de décisions | 65.4% |
| Réseaux de neurones | 59.0% |

[44] H. White - Economic prediction using neural networks : a case of IBM daily stock returns
[45] Using Neural Networks to Forecast Stock Market Prices, 1997
[46] A comparison of Discriminant Analysis versus Artificial Neural Networks, Journal of the Operational Research Society 1993

| | |
|---|---|
| DT + ANN | 77.2% |
| DT + DT | 66.9% |

*Résultats de Tsai et Wang : Comparaison des algorithmes*

## 3.5 Précision des prévisions et rentabilité des stratégies

### 3.5.1 Différentes mesures de performance

On peut distinguer trois grands types de mesures de performances utilisées dans ces études. Les premières sont directement relatives aux modèles. a) Pour les régressions : p-value, t-test pour mesurer la significativité des coefficients et coefficient d'ajustement, coefficient d'ajustement ajusté, MSE pour mesurer la capacité globale de la régression. On retrouve aussi les courbes ROC et RAROC, en particulier dans les études visant à comparer différents algorithmes.

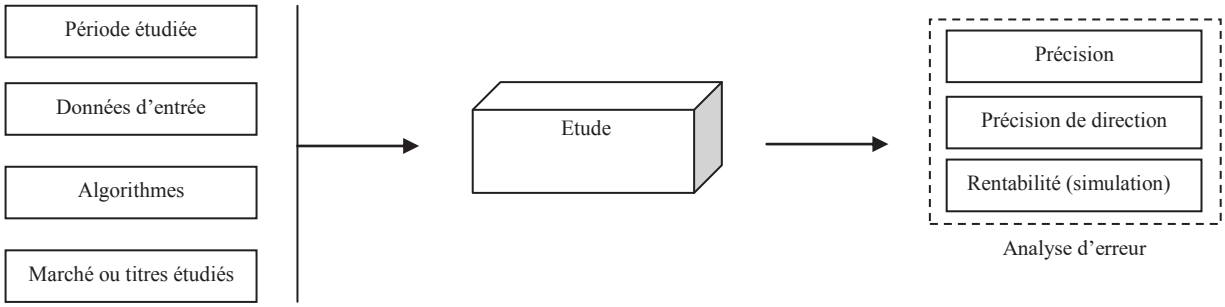
La plupart des études visent à classifier la direction des cours de manière binaire : « haussier » ou « baissier ». La mesure de performance des modèles la plus facilement interprétable est donc la précision[47] de ces modèles. Il est impératif de dissocier la précision sur les données d'entrainement des données étudiées.

Enfin, une troisième méthode de mesure de performance consiste à définir une stratégie de trading basée sur l'algorithme, et à comparer sa rentabilité à des stratégies aléatoires ou à des stratégies « Buy & Hold ». Ces résultats facilement interprétables présentent cependant le biais de la période étudiée (pour des mesures tels que le ratio de Sharpe) ou encore des frais de transactions[48]. Ce dernier point est d'autant plus important que les stratégies nécessitent pour la plupart des ré arbitrages quotidiens.

[47] Voir partie 1.
[48] Chen & al. Chen, Leung, and Daouk (2003) réalisent des simulations avec 3 niveaux de frais de transaction pour valider leurs modèles.

### 3.5.2 Des résultats contrastés

Si les mesures de performances sont relativement standardisées, les facteurs pouvant impacter la qualité des performances sont nombreux :




*Facteurs influant sur les études*

Il est donc délicat de comparer les performances. J'ai choisi de les présenter par stratégies et types de données d'entrées, mais étant donné les spécificités de chaque étude, il n'est cependant pas possible d'en déduire la supériorité d'une stratégie particulière.

#### 3.5.2.1 Stratégies « techniques »

Plusieurs études de Kim s'appuyant à la fois sur un même jeu de données (études de l'indice coréen KOSPI entre 1989 et 1998 à partir de 12 indices techniques), aboutissent à des précisions de 54.7% avec des algorithmes SVM (Kim, 2003), 60.7% avec des algorithmes hybrides GA+CBR (Kim, 2004), 61.70% avec des algorithmes GA+ANN (Kim and Han, 2000). Ils confirment donc une prédictibilité des cours, sans pour autant étudier leur rentabilité. Phua, Zhu, and Koh (2003) trouvent des résultats similaires, avec en moyenne une prédictibilité de 60,7% (jusqu'à 74% suivant les indices étudiées) sur les indices du marché américain. Wang and Chan (2006) utilise un indicateur technique pour prévoir la hausse de 15 ou 20% d'un titre à 100 jours. Il observe une amélioration (de 10,17 à 21,85% suivant les titres) de la précision en utilisant son algorithme par rapport à une stratégie aléatoire d'achat / vente (qui se traduit par un rendement anormal entre 2,4 et 4,7%). Globalement, chacune des études arrive à des niveaux de précisions remettant en cause l'HEM. Il serait cependant intéressant de simuler des stratégies de trading pour chacune d'entre elles afin de voir dans quelle mesure il est possible de profiter de ces retours anormaux (en prenant en compte les frais de transaction).

### 3.5.2.2 Stratégies fondamentales

En 1990, Kimoto et al. (1990) obtiennent une meilleure performance qu'une stratégie B&H sur un indice japonais.[49] Dans leur étude, Tsai et Wang trouvent jusqu'à 77,2% de prédictions correctes à l'aide d'algorithmes hybrides (DT + ANN). Ce résultat est d'autant plus important qu'ils utilisent des arbres de décisions qui leur permettent ainsi de cerner les 12 variables les plus importantes parmi les 53 initiales, et ainsi de pouvoir « expliquer » leur modèle. Toujours en utilisant des réseaux de neurones, mais sur le marché canadien cette fois-ci, Olson and Mossman (2003) obtiennent une précision de 58%.

Une approche similaire sur le marché américain [50] permet à O'Connor and Madden (2006) d'obtenir une rentabilité annuelle de 23,5% par rapport à 13,03% pour l'indice à partir de facteurs externes (e.g. prix de matières premières, taux de changes). Les résultats sont plus contrastés pour Enke and Thawornwong (2005) sur le S&P500 qui utilisent 31 indices économiques et obtiennent une rentabilité annuelle de 22,7% contre 20,1% pour une stratégie buy & hold. Parmi les résultats particulièrement frappants, les différentes stratégies proposées par Chen, Leung, and Daouk (2003) annoncent des rentabilités de plus de 200% contre moins de 10% pour des stratégies B&H, et ce, en considérant des frais de transactions de 3%.

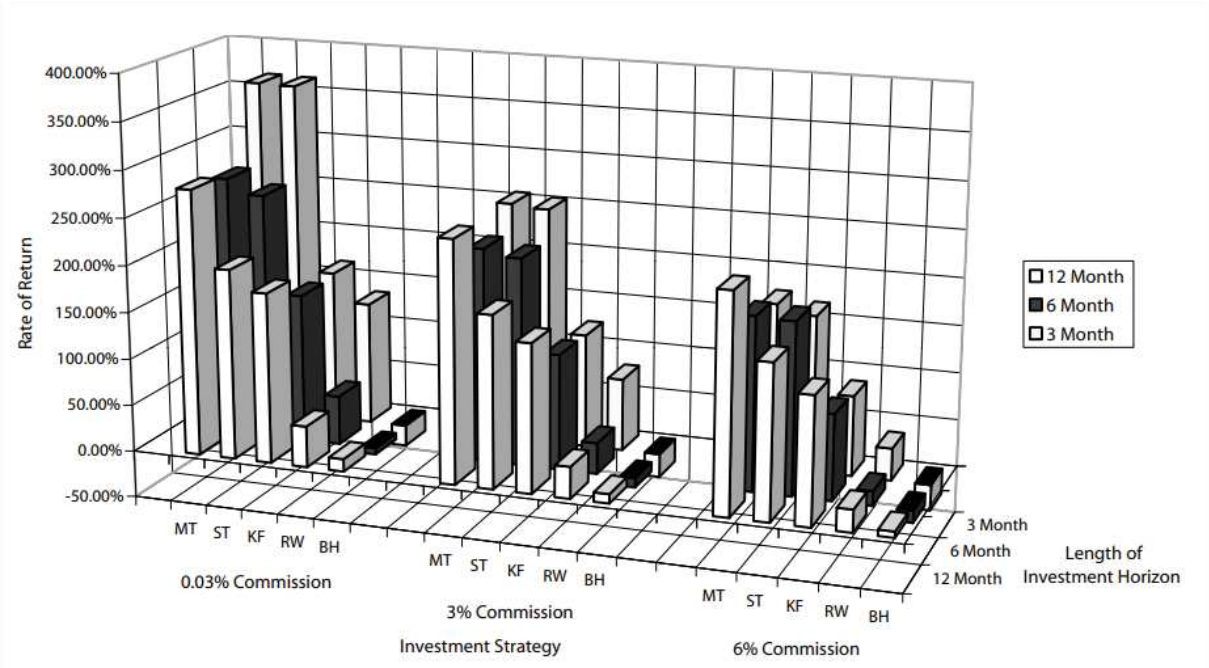



Fig. 3. Annual rate of return for each trading strategy with 3-, 6- and 12-month investment horizons.

*Extrait de l'étude de* Chen, Leung, and Daouk (2003)

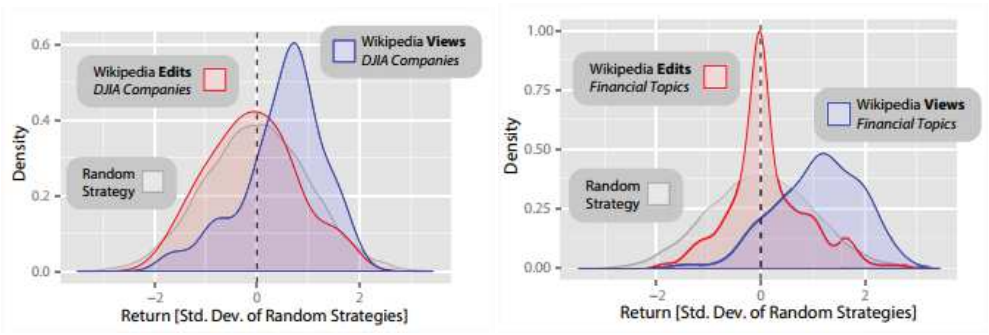
[49] prix init. 1584, 2,5 ans plus tard : 2642 avec stratégie B&H, 3129 avec ANN
[50] Dow Jones Industrial Average Index

Cependant, certaines études basées sur des stratégies fondamentales montrent des résultats moins encourageants (coefficient de détermination de 0,4 pour Datta Chaudhuri, Ghosh, and Eram (2015) sur le marché indien en utilisant des régressions SVM.)

#### 3.5.2.3 Stratégies comportementales et hybrides

Comme nous l'avons évoqué précédemment, les techniques visant à mesurer le nombre d'apparition de mots clefs permettent d'observer une corrélation positive (entre 0,1 et 0,8) entre le volume d'échange et le nombre d'apparitions des mots. C'est le cas de l'étude de Preis & al qui montrent une corrélation (d'au moins 0,3) entre les volumes de recherche d'une entreprise du S&P 500 et son volume de transaction.

Moat et al. (2013) construisent un modèle basé sur la fréquentation des pages Wikipédia relatives aux titres du Dow Jones Industrial Average et aux pages relatives à la finance en général. Leur stratégie s'avère plus performante qu'une stratégie aléatoire (voir graphique-ci-dessous), bien que les résultats varient d'une année à l'autre, en particulier sur 2008.




*Stratégie basée sur les vues Wikipedia (bleu) par rapport à une stratégie aléatoire (grise)*

Les résultats sont partagés lorsque les chercheurs s'appuient sur les articles spécialisés ou les messages issus des réseaux sociaux. Selon Nardo & al (2016) , cela s'explique à la fois par une classification binaire trop simpliste (nécessité d'une classification plus fine) , mais aussi par un effet asymétrique entre le vocabulaire utilisé pour décrire un sentiment haussier ou baissier (beaucoup de mots à tendance baissière n'ayant pas leur équivalent en tendance haussière), ce qui introduit un biais dans les algorithmes. Enfin, il reste la problématique d'un dictionnaire spécifique, d'autant plus importante que les textes analysés sur les réseaux sociaux sont courts (Twitter : 140 caractères).

Tumarkin and Whitelaw (2001) analysent près de 200 000 messages sur Ragging Bull, mais n'observent que des corrélations entre l'humeur et l'évolution de titres uniquement les jours

d'évènements. En revanche, ils montrent une bonne prédictibilité du nombre de messages en fonction des volumes d'échanges de la veille. Das and Chen ne trouvent pas non plus de lien économique entre les sentiments extraits et les rentabilités de titres du S&P 500 à partir d'articles de presse spécialisés et de messages, tout comme Antweiler and Frank (2004) (qui cependant trouvent des corrélations entres nombres d'articles et volumes d'échange), Koppel and Shtrimberg (2006). Tetlock (2007) trouve quant à lui qu'un pessimisme important permet de prévoir une baisse des marchés.

Concernant Twitter, les résultats sont variés. Zhang, Fuehres, and Gloor (2011) observent une faible corrélation négative entre l'humeur de Twitter et les indices américains. Bollen, Mao, and Zeng (2011) trouvent quant à eux une précision dans leurs prédictions de 87,6%, mais ce résultat est à relativiser par l'étude d'une période très spécifique (février à décembre 2008). Karabulut (2013) trouve une corrélation significative entre l'indice de « bonheur » issu de Facebook et la rentabilité des marchés.

Schumaker & al. développent un outil spécifique d'analyse de texte[51] et obtiennent une précision de 58,2%. Ils mettent aussi en avant la nécessité d'intégrer les prix historiques (rentabilité de 2,06% du modèle les intégrant vs 0,6% pour celui n'intégrant que les nouvelles) et préconisent à l'avenir de se concentrer sur des secteurs spécifiques (leur étude portant sur 484 titres du S&P500).

Dans un « Working Paper » de 2017, T. Renault propose alors une régression permettant d'estimer la rentabilité de la dernière demi heure en fonction de 4 paramètres : **1**)la variation d' « humeur » (score entre -1 et 1) entre la veille et la première demi-heure (calculée à partir des algorithmes vus précédemment) **2**) la rentabilité de la 1ere demi-heure **3**) la rentabilité de la 12eme demi heure (issu des travaux de Sun, Najand, and Shen ,2016) **4**) la rentabilité de la dernière demi-heure de la veille.

En utilisant l'algorithme de classification ou le lexique pondéré, le coefficient lié à la variation d'humeur s'avère significatif (à un niveau de 1 % et 0,1% respectivement). La régression obtient ainsi un coefficient de détermination [52] de 2,13 avec le lexique, 2,04 avec l'algorithme de classification. T. Renault montre qu'une stratégie basée sur le sentiment des traders « novices » permet d'obtenir une meilleure rentabilité ajustée que les stratégies

[51] A partir d'AZFin, logiciel d'analyse de texte

[52] Le coefficient de détermination représente la part de la dispersion expliquée par le modèle. Il peut être « ajusté » au nombre de variables explicatives.

traditionnelles. En effet, en appliquant la stratégie qui consiste à acheter ou vendre sur la dernière demi-heure, il obtient une rentabilité de 4,55% avec un ratio de Sharpe de 1,496 (comme le fait remarquer l'auteur, il s'agit d'une performance pour une détention de titre uniquement une demi-heure par jour) et superforme des stratégies basées sur le hasard. Cette étude est particulièrement intéressante car d'actualité et basée sur un modèle relativement simple de régression linéaire.

## 3.6 Limites des modèles prédictifs

### 3.6.1 Limites de certains algorithmes

Les algorithmes sont basés sur des études de corrélation, et non pas de causalité. Des algorithmes comme les réseaux de neurones présentent les limites de la boite noire. Il n'est en effet pas possible pour l'utilisateur de connaitre le poids de chaque variable ni les étapes intermédiaires de l'algorithme. De plus, ce sont des algorithmes plus sensibles à la problématique d' « Overfitting » évoqué précédemment, qui s'explique principalement par des réseaux trop profonds ou par des périodes d'entrainement trop longues. Enfin, ils perdent en performance lorsque le nombre de variables d'entrées est trop important, d'où l'importance de pratiquer des techniques de sélection et de réduction (Yoo, H. Kim, and Jan) que ce soit par des analyses en composantes principales ou par des arbres de décisions (Tsai and Wang).

### 3.6.2 Cout des stratégies

La première question qui se pose dans ces stratégies concerne les couts de transactions, souvent négligés alors que la plupart des stratégies nécessitent des ré arbitrages quotidiens. Par ailleurs, l'analyse des données représente un coût non négligeable pour les entreprises. Bien que la plupart des acteurs financiers aient accès à des sources d'informations (Bloomberg, Reuters), le coût des données est non négligeable (par exemple les solutions GNIP qui permettent aujourd'hui d'accéder à l'ensemble des tweets publiés) (Nardo e al., 2016) De plus, beaucoup de ces données doivent être contrôlées par des humains, en particulier dans les approches basées sur des dictionnaires spécifiques. Enfin, nous avons vu que l'analyse de données est aujourd'hui principalement utilisée pour des prévisions à court terme basées sur des informations récentes, parfois du jour même (Renault, 2017). Il est donc nécessaire d'acquérir une solution IT complète qui permette de gérer ces quantités importantes de données en direct (Yoo, H. Kim, and Jan.).

### 3.6.3 Des anomalies durables ?

Mclean and Pontiff (2016) ont cherché à mesurer l'effet des publications sur la prédictibilité et in fine la rentabilité des portefeuilles constitués. Pour cela ils ont à la fois mesuré la rentabilité des stratégies proposées hors des périodes d'études, puis les rentabilités sur les périodes post-publications. A partir de 97 relations prédictives établies sur des périodes précises (d'en moyenne 44 mois) dans 79 études[53], ils montrent que les rentabilités sont 26% inférieures à celles annoncées par les études hors de ces périodes spécifiques et 58% inférieures sur les périodes post-publication. La différence de 32% s'explique selon les auteurs par l'apprentissage des investisseurs suite aux publications et donc à la disparition de ces « anomalies ». Ils constatent aussi que cette diminution de performance est plus marquée lorsque les rentabilités « anormales » observées dans les études sont plus importantes. Malgré des résultats mitigés et cette problématique de « durabilité », de nombreuses entreprises se spécialisent dans l'analyse de tendance à partir du web pour des usages en finance[54].

## 3.7 Pricing des produits dérivés

La question de la prédictibilité des cours permet de remettre en cause l'hypothèse d'efficience des marchés, et évidemment de déterminer des stratégies plus performantes, profitant de ces anomalies. Cependant, l'enjeu de la prédictibilité s'étend à d'autres paramètres que le prix, et en particulier à la volatilité des titres. C'est une donnée majeure à la base des modèles de pricing des produits dérivés (Hull, 2013). Bien que plus spécifiques que l'étude des cours, de nombreux travaux ont été réalisés sur ce sujet.

Une première série d'études concerne la comparaison des réseaux de neurones avec le modèle de Black-Scholes. Yao & al étudient le pricing des options de l'indice Nikkei 225 (sur une période de 12 mois en 1995) en utilisant un réseau de neurones alimenté par les prix passés (sous entendu que le réseau de neurones « déduira » la volatilité). Pour des marchés volatils (qui remettent en cause l'hypothèse de volatilité constante de Black-Scholes), le réseau de neurones s'avère plus performant pour prévoir les prix de marchés que les méthodes traditionnelles de Black-Scholes. Le modèle s'avère aussi plus performant pour les options américaines (Yao, Li, and Tan, 2000). Ces résultats sont confirmés par Amilon (2003) sur une

[53] Pour une liste complète : https://www2.bc.edu/jeffrey-pontiff/Documents/McLean%20Pontiff%20JF-2015%20Appendix.pdf

[54] Entre autres : https://www.stockpulse.de/en/, http://www.downsidehedge.com/twitter-indicators/

étude d'options sur l'indice suédois entre les années 1997 et 1999. et par Bennell and Sutcliffe (2005) sur les options européennes du TSE 100 index.

D'autres chercheurs se sont intéressés à la prédictibilité de la volatilité. Hamid and Iqbal (2004) étudient la prédictibilité de la volatilité du S&P 500 à partir des prix de 16 autres futures (dont des contrats sur matières premières, des obligations et des devises) mais aussi en fonction des prix spot de 3 indices (DJIA, NYSE Composite Index et évidemment le S&P 500) sur une période de 10 ans entre 1984 et 1994. Ils comparent leurs résultats à la volatilité implicite issue du modèle de pricing BAW (Barone-Adesi and Whaley Mode, basé sur le modèle de Black-Scholes). Les résultats montrent une meilleure performance des réseaux de neurones pour estimer les volatilités sur 15 jours et 35 jours. Enfin, des études plus récentes proposent des modèles hybrides utilisant à la fois des réseaux de neurones et les modèles GARCH qui permettent d'atteindre de meilleures performances dans le pricing des produits dérivés que les modèles GARCH seuls (Wang, 2009) (Kristjanpoller, Fadic, and Minutolo (2014).

Ces études utilisent principalement en variables d'entrées des données quantitatives à l'image des modèles de pricing traditionnels. Il serait intéressant d'intégrer dans ces réseaux de neurones des mesures de sentiments issus des réseaux sociaux (Nardo et al., 2016).

# Conclusion

Cette étude met en avant le rôle grandissant du Machine Learning en Finance. Bien que les algorithmes utilisés ne soient pas nouveaux, les capacités de calcul et les nouvelles données disponibles permettent aujourd'hui d'améliorer les précisions des analyses prédictives. Cela devient un avantage compétitif dans des domaines variés comme la tarification, le risque de défaillance, la détection de fraude ainsi que la gestion de portefeuille. A ce titre, il est primordial pour les acteurs de la Finance de comprendre les principes et les applications et du Machine Learning.

La question de la prédictibilité des cours en utilisant les nouveaux algorithmes de Machine Learning revient fréquemment dans les articles académiques. L'étude de 42 études m'a permis de mieux comprendre les tendances de la recherche dans ce domaine. Les algorithmes utilisés sont principalement des algorithmes de classifications visant à prévoir une hausse ou une baisse des titres. Les réseaux de neurones et leurs dérivés sont les plus présents dans la littérature, mais les modèles hybrides sont plus performants, et permettent dans certains cas de palier au problème de la « boite noire » rencontré avec les réseaux de neurones. Concernant le développement de nouveaux algorithmes, le Meta-Learning, qui vise à optimiser le choix des algorithmes est un domaine prometteur. Cependant, des études récentes utilisant des régressions multiples obtiennent aussi de très bons résultats, ce qui confirme l'importance de la qualité des données par rapport à l'algorithme. Ces données ont beaucoup évolué durant les 30 dernières années, en particulier avec l'analyse de texte, et plus récemment les réseaux sociaux. Il s'agit là aussi d'un champ de recherche récent, dont les premiers résultats semblent mitigés et confirment la nécessité d'intégrer des connaissances métiers pour entrainer efficacement les modèles.

Etant donné d'une part la spécificité de chaque étude, d'autre part mon manque de connaissances sur certains points spécifiques liés à la mesure de performance, il serait hasardeux de tirer des conclusions radicales concernant l'évolution des performances grâce aux nouveaux algorithmes de Machine Learning. On peut cependant retenir que les précisions obtenues (de l'ordre de 65% lorsqu'il s'agit de prévoir la direction des cours) ne progressent pas de manière significative sur ces 25 dernières années. Comme le remarquent Mclean and Pontiff (2016), les publications ont un impact direct sur la rentabilité des stratégies proposées.

Cela pose donc la question de la rentabilité «nette » de telles stratégies : frais d'acquisition des données, frais de développement, frais de transactions et durabilité de la stratégie.

Ce travail de recherche soulève plusieurs pistes pour des études futures. D'une part, il est essentiel de construire des lexiques spécifiques au domaine de la finance afin de mieux mesurer les sentiments issus du web. En poussant ce concept plus loin, il serait intéressant de se concentrer sur un secteur d'activité, voir un titre en particulier en étudiant des données spécifiques (par exemple les immatriculations de voitures pour le domaine automobile, ainsi que des commentaires de forums spécialisés). Une autre piste de recherche proposée par Nardo & al (2016) concerne l'étude des corrélations entre volatilité des sentiments extraits du web et volatilité des titres.

De manière plus générale, il me semble que dans des domaines aussi liés à l'informatique, le support traditionnel des articles de recherche pourrait être enrichi par des plateformes collaboratives (à l'image des plateformes de partage de code), mettant à la fois à disposition des données et les outils utilisés pour chaque étude. Cela permettrait de rendre les études plus dynamiques et plus facilement comparables (en réutilisant par exemple les algorithmes d'études précédentes sur des jeux de données actuelles).

# Bibliographie


Abbasi, Ahmed, Conan Albrecht, Anthony Vance, and James Hansen, 2012, Metafraud: a meta-learning framework for detecting financial fraud, *Mis Quarterly* 36, 1293–1327.

Amilon, Henrik, 2003, A neural network versus Black-Scholes: a comparison of pricing and hedging performances, *Journal of Forecasting* 22, 317–335.

Antweiler, and Frank, 2004, Is all that talk just noise ? The information content of Internet stock message boards, .

Avery, Christopher N., Judith A. Chevalier, and Richard J. Zeckhauser, 2016, The "CAPS" prediction system and stock market returns, *Review of Finance* 20, 1363–1381.

Baesens, Bart, Rudy Setiono, Christophe Mues, and Jan Vanthienen, 2003, Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, *Management Science* 49, 312–329.

Baker, Malcolm, and Jeffrey Wurgler, 2007, Investor sentiment in the stock market, *The Journal of Economic Perspectives* 21, 129–151.

Ben Jabeur, Sami, and Youssef Fahmi, 2013, Prévision de la défaillance des entreprises : une approche de classification par les méthodes de Data-Mining, *Gestion 2000* 31, 31.

Bennell, Julia, and Charles Sutcliffe, 2005, Black–Scholes versus artificial neural networks in pricing FTSE 100 options, .

Bollen, Johan, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of computational science* 2, 1–8.

Boyer, Jean-Marc, 2015, La tarification et le *big data* : quelles opportunités ?, *Revue d'économie financière* 120, 81.

Brazdil, P., Carrier, Soares, and Vilalta, 2008, *Metalearning : Applications to Data Mining*.

Changa, Ying-Hua, and Shih-Chin Wangb, 2013, Integration of Evolutionary Computing and Equity Valuation Models to Forecast Stock Values Based on Data Mining, *Asia Pacific Management Review* 18, 63–78+.

Chen, An-Sing, Mark T. Leung, and Hazem Daouk, 2003, Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index, *Computers & Operations Research* 30, 901–923.

Das, S.R., and M.Y. Chen, Yahoo! For amazon : sentiment extraction from small talk on the web, *Journal of Economic Perspectives*.

Datta Chaudhuri, Tamal, Indranil Ghosh, and Shahira Eram, 2015, Predicting stock returns of mid cap firms in India–an application of random forest and dynamic evolving neural fuzzy inference system, .

Desrosières, Alain, 2010, *La Politique Des Grands Nombres, Histire de La Raison Statistique*.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg, 2012, How are shorts informed?: Short sellers, news, and information processing, *Journal of Financial Economics* 105, 260–278.

Enke, David, and Suraphan Thawornwong, 2005, The use of data mining and neural networks for forecasting stock market returns, *Expert Systems with Applications* 29, 927–940.

Fallahi, Saeed, Meysam Shaverdi, and Vahab Bashiri, 2014, Applying GMDH-type neural network and genetic algorithm for stock price prediction of Iranian cement sector, *Soft-Computing in Capital Market: Research and Methods of Computational Finance for Measuring Risk of Financial Instruments*, 147.

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *The Journal of Finance* 25, 383.

Gandomi, Amir, and Murtaza Haider, 2015, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management* 35, 137–144.

Gensai, Ramazan, 1996, Non-linear prediction of security returns with moving average rules.pdf, *Journal of Forecasting*.

Golmohammadi, Koosha, and Osmar R. Zaiane, 2012, Data Mining Applications for Fraud Detection in Securities Market, (IEEE).

Hamid, Shaikh A., and Zahid Iqbal, 2004, Using neural networks for forecasting volatility of S&P 500 Index futures prices, *Journal of Business Research* 57, 1116–1125.

Hong, Taeho, and Ingoo Han, 2004, Integrated approach of cognitive maps and neural networks using qualitative information on the World Wide Web: the KBNMiner, *Expert Systems* 21, 243–252.

Hull, John C., 2013, *Options, Futures, and Other Derivatives*.

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, 2012, Big data: The Next Frontier for Innovation, Competition, and Productivity, *McKinsey Global Institute*.

Karabulut, Yigitcan, 2013, Can Facebook predict stock market activity?, .

Kearney, Colm, and Sha Liu, 2014, Textual sentiment in finance: A survey of methods and models, *International Review of Financial Analysis* 33, 171–185.

Keim, Daniel A., 2002, Information Visualization and Visual Data Mining, *IEEE Transactions on visualization and computer graphics*.

Khashman, Adnan, 2010, Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes, *Expert Systems with Applications* 37, 6233–6239.

Kim, and Kim, 2014, Investor Sentiment from Internet Message Postings and Predictability of stock returns, .

Kim, Kyoung-jae, 2003, Financial time series forecasting using support vector machines, *Neurocomputing* 55, 307–319.

Kim, Kyoung-Jae, 2004, Toward global optimization of case-based reasoning systems for financial forecasting, *Applied intelligence* 21, 239–249.

Kim, Kyoung-jae, and Ingoo Han, 2000, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, *Expert systems with Applications* 19, 125–132.

Kimoto, Takashi, Kazuo Asakawa, Morio Yoda, and Masakazu Takeoka, 1990, Stock market prediction system with modular neural networks, *Neural Networks, 1990., 1990 IJCNN International Joint Conference on* (IEEE).

Kohara, K., Stock price prediction Using Prior Knowledge and Neural Networks.pdf, *Intelligent system in Accounting, Finance and Management*, 1997.

Koppel, Moshe, and Itai Shtrimberg, 2006, Good news or bad news? let the market decide, *Computing attitude and affect in text: Theory and applications*, 297–301.

Kristjanpoller, Werner, Anton Fadic, and Marcel C. Minutolo, 2014, Volatility forecast using hybrid Neural Network models, *Expert Systems with Applications* 41, 2437–2442.

Lawrence, Ramon, 1997, Using neural networks to forecast stock market prices, *University of Manitoba*.

Leung, Henry, and Thai Ton, 2015, The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks, *Journal of Banking & Finance* 55, 37–55.

Mclean, R. David, and Jeffrey Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability?: Does Academic Research Destroy Stock Return Predictability?, *The Journal of Finance* 71, 5–32.

Menkhoff, Lucas, 1997, Examining the use of technical currency analysis.pdf, .

Moat, Helen Susannah, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis, 2013, Quantifying Wikipedia Usage Patterns Before Stock Market Moves, *Scientific Reports* 3.

Nardo, Michela, Marco Petracco-Giudici, and Minás Naltsidis, 2016, WALKING DOWN WALL STREET WITH A TABLET: A SURVEY OF STOCK MARKET PREDICTIONS USING THE WEB: CAN THE WEB PREDICT THE STOCK MARKET, *Journal of Economic Surveys* 30, 356–369.

Ng, Keng-Hoong, and Kok-Chin Khor, 2017, StockProF: a stock profiling framework using data mining approaches, *Information Systems and e-Business Management* 15, 139–158.

Ngai, E.W.T., Yong Hu, Y.H. Wong, Yijun Chen, and Xin Sun, 2011, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision Support Systems* 50, 559–569.

O'Connor, Niall, and Michael G. Madden, 2006, A neural network approach to predicting stock exchange movements using external factors, *Knowledge-Based Systems* 19, 371–378.

Olson, Dennis, and Charles Mossman, 2003, Neural network forecasts of Canadian stock returns using accounting ratios, *International Journal of Forecasting* 19, 453–465.

Phua, Paul Kang Hoh, Xiaotian Zhu, and Chung Haur Koh, 2003, Forecasting stock index increments using neural networks with trust region methods, *Neural Networks, 2003. Proceedings of the International Joint Conference on* (IEEE).

Renault, Thomas, 2017, Intraday online investor sentiment and return patterns in the U.S. stock market, .

Schumaker, Robert P., and Hsinchun Chen, Textual Analysis of Stock Market Prediction Using breaking Financial News : The AZFinText System, .

Sun, Licheng, Mohammad Najand, and Jiancheng Shen, 2016, Stock return predictability and investor sentiment: A high-frequency perspective, *Journal of Banking & Finance* 73, 147–164.

Swan, Melanie, 2013, The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery, *Big Data* 1, 85–99.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *The Journal of Finance* 62, 1139–1168.

Travaille, Peter, Roland M. Müller, Dallas Thornton, and Jos Hillegersberg, 2011, Electronic fraud detection in the US medicaid healthcare program: lessons learned from other industries, .

Tsai, C-F., and Wang, S-P., Stock Price Forecasting by Hybrid Machine Learning Techniques, .

Tumarkin, Robert, and Robert F. Whitelaw, 2001, News or noise ? Internet Message Board Activity and Stock Prices, .

Viaene, Stijn, Richard A. Derrig, Bart Baesens, and Guido Dedene, 2002, A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, *Journal of Risk and Insurance* 69, 373–421.

Wang, J, and S Chan, 2006, Stock market trading rule discovery using two-layer bias decision tree, *Expert Systems with Applications* 30, 605–611.

Wang, Shiguo, 2010, A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research, (IEEE).

Wang, Yi-Hsien, 2009, Nonlinear neural network forecasting model for stock index option price: Hybrid GJR?GARCH approach, *Expert Systems with Applications* 36, 564–570.

White, Halbert, 1988, Economic prediction using neural networks: The case of IBM daily stock returns, .

Yao, Jingtao, Yili Li, and Chew Lim Tan, 2000, Option price forecasting using neural networks, *Omega* 28, 455–466.

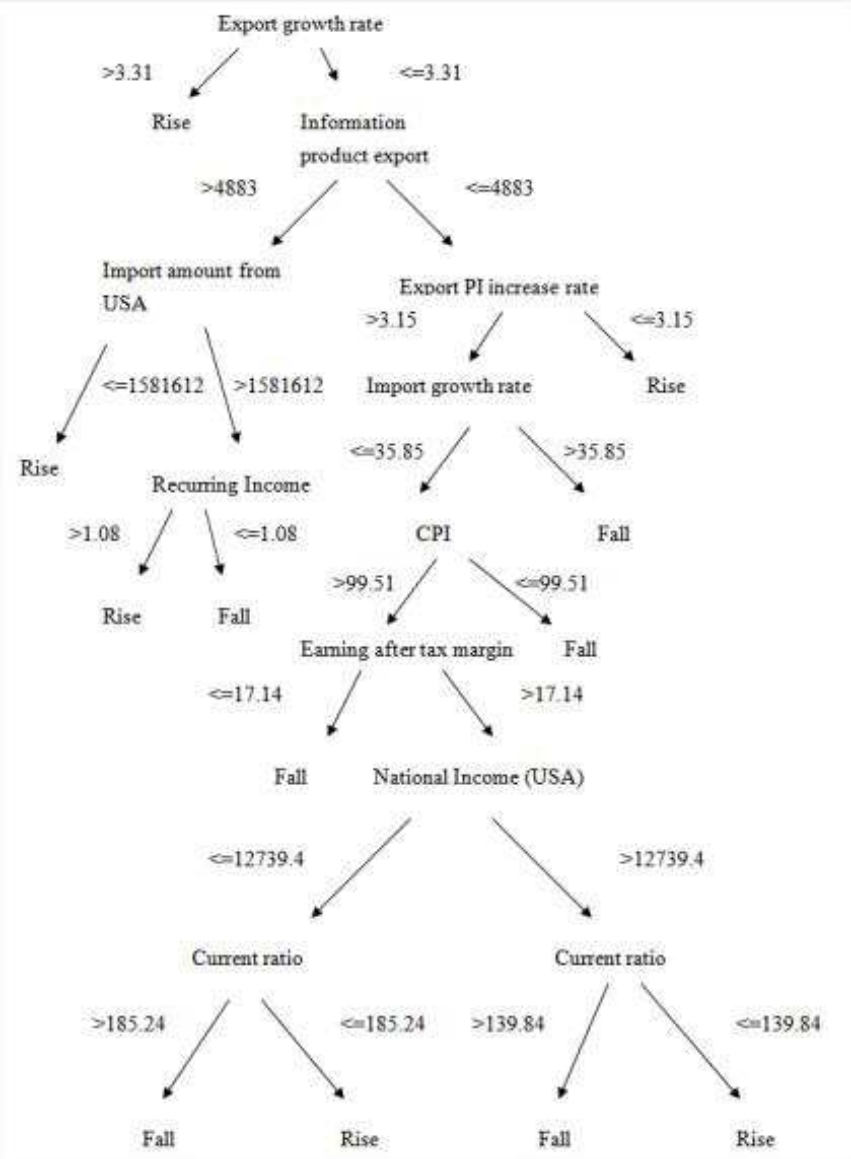
Yoo, Paul D., Maria H. Kim, and Tony Jan, Mahine Learning techniques and Use of Event Information for Stock Market Prediction : a survey and evaluation, .

Yoo, Paul D., Maria H. Kim, and Tony Jan, 2005, Machine learning techniques and use of event information for stock market prediction: A survey and evaluation, *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on* (IEEE).

Yue, Dianmin, Xiaodan Wu, Yunfeng Wang, Yue Li, and Chao-Hsien Chu, 2007, A review of data mining-based financial fraud detection research, *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on* (Ieee).

Zhang, Xue, Hauke Fuehres, and Peter A. Gloor, 2011, Predicting Stock Market Indicators Through Twitter ?I hope it is not as bad as I fear?, *Procedia - Social and Behavioral Sciences* 26, 55–62.

# Annexe A




*Exemple d'arbre de décision tiré de l'étude de Tsai et Wang*

# Annexe B : Etudes sur la prédictibilité des cours

| date | Titre de l'article | type de valeur étudiée | Zone géo. | Données d'entrée | date début obs. | Durée d'étude (années) | Algorithmes utilisés | Résultat/Précisions |
|---|---|---|---|---|---|---|---|---|
| 2017 | Intraday online investor sentiment and return patterns in the US stock market. | Indice | Amérique du nord | StockTwits | | 0,00 | lexique pondéré, lexique spécifique, maximum entropy classifier + régression multiple. | surperfomance en utilisant la stratégie visant à étudier le sentiment de la première demi-heure et à acheter/vendre sur la dernière demi-heure. (ratio de sharpe de 1,497). |
| 2017 | StockProF: a stock profiling framework using data mining approaches | Titres individuels | Asie | ratios financiers (issus de Datastream, Reuters) : total asset turnover / cash ratio / debt ratio / return on equity / dividend yield / Price earning ratios | 01/01/2013 | 1,00 | Local Outlier Factor + Expectation Maximization (clustering) = (StockProf) | 3 catégories : surperformants et agressifs / défensifs en cas de crises ou de ralentissement du marché / Investissements à éviter |
| 2016 | Stock return predictability and investor ssentiment : a high-frequency perspective. | Indice | Amérique du nord | MArketPsych Indices (TRMI) - https://www.marketpsych.com/data/ -lagged half hur investor sentiment + lagged macroeconomic variables | 01/01/1998 | 14,01 | régression : fonction de la variation de sentiment dans la demi heure précédente. 2eme modèle prenant en compte les rentabilités de l'instant t et de la demi-heure précédente | régression de plus en plus précise que les heures étudiées sont proches de la fin de journée dernières heures) (Adj. R2 jusqu'à 3,66) . Sentiment plus important que l'effet momentum. Plus important en été ( et durant les périodes de croissances (jusquà 2,04). Variables macro économiques non explicatives |
| 2016 | The "CAPS" prediction system and stock market returns | | Amérique du nord | 5 million stock "picks" (indication de tendance relative à un titre sur une période donnée) www.caps.fool.com générés par 60 000 pers. | 01/11/2006 | 5,17 | simulation d'achat et de vente suivant les picks. | renta annulaisée 12% - les picks négatifs sont plus influents. Performance s'explique par le choix des titres et non par le market timing. |
| 2017 | Intraday momentum : the first half-hour return predicts the last half-hour return | Indice | Amérique du nord | rentabilité sur la première demi-heure (par rapport à la veille) -Trade and Quote Database | 01/01/1993 | 21,01 | régression : rentabiltié de la dernière demi-heure fonction de la rentabilité de la première demi-heure. | coefficients significatifs (R2 1,6) - surtout dans les jours plus volatils (R2 3,3), ou avec plus d'échange, sur les périodes de récession et les périodes d'évènements. |
| 2015 | Predicting stock returns of mid cap firms in India-an application of random forest and dynamic evolving neural fuzzy inference system | Titres individuels | Asie | ratios financiers : P/E, Price to Book Debt Equity Rati, Inteerest coverage Ratio, Gross Profit Margin, Divident Payout Ratio Extent of Promoter Holding of Shares, Sectoral Returns, Volume. | 01/01/2013 | 3,00 | unsupervised feature selection (Analyse en composante principales et dérivées) / random forest, bagging (bootstrap aggreagating, similaire aux RF), boosting and support vector regression puis modèle hybride . Pour réaliser ce moodèle hybride, ils utilisent un algorithme "Evolutionary Algorithm" (équivalent d'une sélection naturelle) | R2 globalement basses. (max 0,4) Meilleures prévisions pour les large cap et mid-cap que les small-cap. Modèle hybride plus performant. |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2015 | The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks | Titres individuels | Asie | 2,5 millions HotCopper messages (accompagnés d'un label parmi 7 sentiments) | 01/01/2003 | 6,00 | Naive Bayesian pour l'analyse de sentiment. Puis régressions multiples. | Corrélation entre volume de message et volume d'échange de titres pour les small cap. Corrélation entre les sentiments et les rentabiltiés à deux jours. |
| 2014 | Investor sentiment from Internet message postings and the predicatbility of stock returns. | Titres individuels | | 32 million de messages postés sur Yahoo Finance Message board avec labels :"strong buy, buy, hold, sell, strong sell". Mais aussi caractéristiques de l'auteur. | 01/01/2005 | 6,00 | Naive Bayesian pour l'analyse de sentiment ((http://www.nltk.org)). Puis régressions multiples. | Pas de corrélation entre les sentiments et la rentabilité des titres (individuels ou agrégés sous des indices), ni sur leur volatilité et leur volume d'échange. En revanche, influence des prix passés sur le sentiment des investisseurs. |
| 2014 | Applying GMDH type neural network and genetic algorithm for stock price prediction of Iranian cement sector | Titres individuels | Asie | ratios : Earning per shapres, Prediction Earnings Per Share, Dividend per Share , Price earning ratio. | 01/01/1999 | 10,01 | Group Methode of Data Handling (GMDH) (réseau de neurones). | |
| 2013 | Can Facebook predict stock market activity? | Indice | Amérique du nord | Reuteurs Datastream / Facebook's Gross National Happiness | 01/01/2008 | 5,00 | vector autoregressive (VAR) | Corrélation entre l'indice de "bonheur" et la rentabilité des marchés. (t -stat 2,54) en prenant l'humeur un jour auparavant. Coefficient de 12,1024 : augmentation de 12 pt de bases pour un une augmentation d'un écart type. |
| 2013 | Quantifying Wikipedia usage patterns before stock market moves | Indice | Amérique du nord | pages vues (stats.grok.se) et éditées sur wikipedia | 01/01/2007 | 6,00 | stratégie d'achat / vente suivant l'évolution (increase / decrease) des pages vues / éditées. Comparée à une stratégie aléatoire. | stratégies basées sur les vues des pages relatives aux titres du Dow Jones surperforment. (voir distribution) en revanche, pas celles basées sur les éditions. Dépend fortement des années étudiées (crises 2008 2011) |
| 2013 | Integration of Evolutionnary Computing and equity Valuation Models to Forecast Stock Values Based on Data Mining | Titres individuels | Asie | Rapports financiers, puis modèle d'Ohlson / CEO shareholding ratio / composition de la direction | 01/01/2005 | 5,00 | ANN (classification)+ GA | Classification des entreprises : à la fois dans un but de conseil "interne" mais aussi pour les investisseurs. |
| 2012 | How are short informed ? Short sellers, news and information processing | Titres individuels | Amérique du nord | short-sales (NYSE Trade and Quote Regulation SHO database) + news (Dow Hones News Service stories and Wall Street Journal stories | 01/01/2005 | 3,00 | Dictionnaire Harvard IV, et liste de mot négatifs développée par Loughran et MacDonald. Score = nombre de mots négatifs / nombre de mots de l'article. | Les "short sales" anticipent rarement les informations, mais profitent des informations disponibles. Ils analysent mieux l'information que les autres acteurs. Les "short salers" les mieux informés sont les clients et non les market makers. |
| 2011 | Twitter mood predicts the stock market | Indice | Amérique du nord | Twitter (10 millions par 2,7M d'utilisateurs). | 28/02/2008 | 0,81 | OpinionFinder + Google-Profile of Mood States (calm, alert, sure, vital, kind, happy). Granger causality analysis + self organizing fuzzy neural netwok. | prédiction des tendances haussières ou baissières quotidiennes : 87,6 % en utilisant GPOMS |
| 2010 | Predicting Stock Market Indicators through Twitter ? I hope it is not as bad as I fear ? | Indice | Amérique du nord | Twitter | 30/03/2009 | 0,44 | mesure du nombre d'apparition de certains "mots" (tweet / hope / happy / fear / worry / nervous / anxious / Upset / Positive / negative) | faible corrélation négative entre le "sentiment" twitter et les indices. (-0,323 avec NASDAQ) |
| 2009 | Nonlinear neural network forecasting model for stock index option price : Hybrid GJR-GARCH approach | Autres | Asie | 21120 calls - paramètres de B&S | 03/01/2005 | 1,99 | Grey GJR GARCH volatility (approche asymétrique de la volatilité). + ANN | Grey GJR GARCH volatilty présente une meilleure prédictabilité que des approches pus classifques (type GJR GARCH ou B&S) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2007 | Giving content to investor sentiment: The role of media in the stock market | Indice | Amérique du nord | Articles Wall Street Journal | 01/01/1984 | 15,72 | General Inquirer (compte le nombre de mots dans les articles qui appartiennent à certaines catégories). Basée sur le dictionnaire Harvard 4. Analyse en composantes principales. | Pessimisme important permet de prévoir une baisse des marchés suivie d'un rééquilibrage ainsi que des volumes d'échanges importants. (coefficients significatifs à 1%) |
| 2007 | Predicting stock index incrments by neural networks : the role of trading volume under different horizons | Indice | Amérique du nord | Volumes échangés | 01/01/1997 | 9,00 | ANN | Confirme l'hypothèse que les volumes d'échanges permettent de prévoir les mouvements des indices. (uniquement à court terme) |
| 2007 | A hybrid approach bsaed on neural netorks and genetic algorithms for deteting temporal patterns in stock markets | Indice | Asie | valeurs historiques de l'indice | 01/01/1997 | 2,92 | ANN ATNN (Adaptive time delay) TDNN (time delay) GA | Modèles hybrides > Modèle simples |
| 2007 | Forecasting the volatility of stock price index | Indice | Asie | 930 prix d'option (quand ?) pour l'entrainement. | | | ANN, NN-EWMA, NN-GARCH and NN-EGARCH : réseaux de neurones et analyses de séries historiques. | ANN hybrides ont plus de succès |
| 2007 | Stock Market trading rule discovery using two-layer bias decision tree | Titres individuels | Amérique du nord | Un indicateur technique : RSI (relative strenght index) calculé sur deux périodes (14 et 28 jours) | 01/01/1990 | 12,62 | two layer decision tree. Prévisions d'augmentation supérieures à 15% ou 20% à 100 jours. | Bonnes prédictions sur le marché des actions (rentabilité supérieure entre 2,4 et 4,7 % par rapport à un achat aléatoire). |
| 2006 | Good news or bad news ? Let the market decide. | Indice | Amérique du nord | articles concernant les titres du S&P 500 (12 000 - en moy. 24 par titre) - Multex Significant Developments corpus | 01/01/2002 | 1,00 | Deux approches :a) labeliser l'article suivant le mouvement du titre en question. (en comparant le prix à l'ouverture le lendemain par rapport à la fermeture la veille de la publication) b) enfonction des mouvemenst sur les deux jours suivants avec critère de mouvement de 10% (à la hausse ou 7,8% à la baisse). SVM, Naive Bayesian, DT | Le modèle fonctionne avec la première approche mais ne permet pas d'en "profiter". En revanche, il n'est pas significatif avec la deuième approche (uniquement 52% de prédictibilité) d'influence du choix de l'algorithme sur ces résultats. |
| 2006 | Textual analysis of Stock Market Prediction Using breking Financial News : the AZFinText System | Titres individuels | Amérique du nord | 9211 articles financiers et 10,259,000 cotations. articles Wall Street Journall / Bloomberg | 26/10/2005 | 0,09 | Analyse de texte : AZFin (méthode proper noun)<br>Prédiction des cours : régressions sur les prix historiques puis dérivé SVM | Analyse de texte : Proper Noun > Bag of words.<br>Prédiction des cours : MSE : 0,04261 sur la prédiction du prix. Précision de 58,2% pour le sens du prix et rentabilité de 2,84% return |
| 2006 | A neural network approach to predicting stock exchange movements using external factors | Indice | Amérique du nord | facteurs externes : prix de matières premières (WTI Cushing Crude Oil ), taux de changes (OANDA.com) | 02/01/1986 | 19,10 | ANN | meilleure performance que le marché : (23,5% / an par rapport à 13,03% pour l'indice). |
| 2006 | The application of neural networks to fercast fuzzy time series | Indice | Asie | Prix historiques | 01/01/1991 | 13,01 | Back propagation ANN , Chen's model (time series), modèle hybrides | Modèle hybride > ANN >Fuzzy Time series (marches aléatoires). |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2005 | The use of data mining and neural networks for forecasting stock market returns | Indice | Amérique du nord | différences de valeurs (Pt - Pt-1) données mensuelles. Ajustées des saisonnalités. 31 ratios et indices économiques | 01/03/1976 | 23,77 | Selection de 15 variables, normalisation. ANN + regression linéaire | régression et certains ANN génèrent une rentabilité inférieure à l'indice ou à une stratégie aléatoire. Il arrive à obtenir des rentabilités mensuelles de 1,72% contre 1,54% pour une stratégie "buy and hold") |
| 2004 | Toward Global Optimization of Case-based Reasoning Systems for Financial Forecasting | Indice | Asie | 12 indicateurs "techniques" (uniquement basés sur les prix passés) | 01/01/1989 | 9,01 | CBR + GA > CBR | GA + CBR : 60,7% CBR (et variantes) ; entre 52,14 et 54,71% |
| 2004 | Is all that talk just noise? The information content of Internet stock message boards. | Indice | Amérique du nord | Yahoo Finance + Raging Bull sur 45 titres (actions) partagés entre "old economy" et "new economy". + articles wall Street Journal | 01/01/2000 | 1,00 | Classification des messages : Naive Bayes / SVM | Pas de prévision significative avec le modèle bayésien naif. Corrélation entre le nombre d'articles et les volumes d'échange ainsi que la volatilité. Confirmation de l'effet "lundi" |
| 2003 | Financial time series forecasting using support vector machines | Indice | Asie | 12 indicateurs "techniques" (uniquement basés sur les prix passés) | 01/01/1989 | 9,01 | SVM | SVM > NN & CBR |
| 2003 | Forecasting Stock Index Increments Using Neural Networks with Trust Region Methods | Indice | Amérique du nord | données historiques des indices | 04/01/1994 | 8,74 | ANN | succès moyen de plus de60% et meilleures prédictions a 74% en utilisant un modèle "trusted region based" neural network. |
| 2003 | Application of neural neworks to an emerging financial market : foreasting and trading the Taiwan stock Index | Indice | Asie | Prix historiques, taux d'interets, indices de consommation, PIB, niveau de production | 01/01/1982 | 10,59 | PNN (probabilistic Neural Network) -GMM (General Methods of Moments) | PNN ont plus de succès que GMM. Stratégies de trading avec différentes comissions. Même avec 3% de comissions, rentabilités annuelles allant jusqu' à 255% contre 6,54% pour B&H….. |
| 2003 | Neural Network forecasts of Canadian stock returns using accounting ratios | Titres individuels | Amérique du nord | 61 ratios financers pour 2352 entreprises canadiennes. | 01/01/1976 | 17,01 | ANN | ANN > Regressions linéaires Jusqu'à 58% de prédictibilité (direction) |
| 2001 | Yahoo! for Amazon: Sentiment extraction from small talk on the web | Titres individuels | Amérique du nord | 25000 messages classés entre "bullish bearish neutral" | 01/10/2000 | 0,25 | Naive Classifier, SVM, Discriminant based Classifier, Adjective Adverb Phrase Classifier, Bayesien classifier | Pas de lien économique entre les sentiments exraits et la rentabilité du titre, mais bonne performance des algorithmes de classifications pour mesurer le sentiment des investisseurs. |
| 2001 | News or noise ? Internet Message Board Activity and Stock Prices | Titres individuels | Amérique du nord | RaggingBull 181000 messages | 01/04/1999 | 0,88 | Analyse VAR (Vector autoregression) | Corrélations observables entre humeur et évolution des titres uniquement les jours de forte activité sur Internet. Possibilité de prévoir le nombre de messages en fonction des volumes échangés la veille. |
| 2000 | Genetic algorithms approach to feature discretiation in artificial neural netorks for the prediction of stock price index | Indice | Asie | indices "techniques" | 01/01/1989 | 9,01 | ANN + GA | accuracy : 82% dans la prédiction des tendance haussières et baissières hebdomadaires. |
| 1999 | Improving returns on stock invesment through neural network selection | Indice | Amérique du nord | 7 ratios : rendements (P/E) / liquidité / risque / croissances / momentum | 01/01/1993 | 4,00 | ANN qui permet de choisir des titres | Surperformance par rapport à l'index |
| 1997 | Stock Price Prediction Using Prior Knowledge and Neural Networks | Indice | Asie | Presse + indicateurs économiques + prix historiques | 01/08/1989 | 1,58 | ANN, régressions linéaires | ANN > Réressions linéaires (5% significativité) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1996 | Non-linear prediction of security returns with moving average rules | Indice | Amérique du nord | prix historiques journaliers | 01/01/1963 | 26,02 | régressions + ? | Surperformance des modèles, remise en cause de l'hypothès 'efficence des marchés |
| 1988 | Economic prediction using neural networks : a case of IBM daily stock returns | Titres individuels | Amérique du nord | | 01/01/1974 | 5,00 | ANN | |
| 2009 | Stock Price Forecasting by Hybrid Machine Learning Techniques | Indice | Asie | BDD TEJ - 53 variables économiques - sur 5 périodes | 01/01/2002 | 5,42 | Analyse en composantes principales => 53 variables. ANN, DT, ANN + DT, DT + DT | DT : 65,4 % ANN : 59% DT + ANN : 77,2% DT+DT : 66,85 % |
| 1990 | Stock Market prediction system with modular neural networks | Indice | Asie | indices techniques et économiques ( >6) | 09/01/1987 | 2,49 | modular neural networks , régressions multiples | MNN > Régressions multiples. surperformance par rapport à B&H prix init. 1584 prix fin avec modèle : 3129 vs 2642 (B&H). Sur env 2 ans. |