

**MODELES CLASSIQUES ET DE DATAMINING LES PLUS UTILISES
EN EVALUATION ET EN PREDICTION DE LA PERFORMANCE DES
ACTIONS**

-Revue de littérature-

**CLASSIC AND DATAMINING MODELS MOST USED IN
EVALUATION AND PREDICTION OF THE PERFORMANCE OF
STOCKS**

-Literature review-

Abdellilah Nafia

Doctorant à la Faculté des Sciences Juridiques, Économiques et Sociales, Souissi
Université Mohammed V, Rabat, Maroc
Laboratoire d'Analyse Économique et de Modélisation
Email: nafiable@gmail.com

Abdellah Youssefi

Professeur à la Faculté des Sciences Juridiques, Économiques et Sociales, Souissi
Université Mohammed V, Rabat, Maroc
Laboratoire d'Analyse Économique et de Modélisation
Email: a.youssefi@um5s.net.ma

Abdellah Echaoui

Professeur à la Faculté des Sciences Juridiques, Économiques et Sociales, Souissi
Université Mohammed V, Rabat, Maroc
Laboratoire d'Analyse Économique et de Modélisation
Email: a.echaoui@um5s.net.ma

Résumé

Beaucoup de recherches ont été avancées sur la question sur la prévisibilité de la performance des titres. Depuis longtemps, les investisseurs cherchent à trouver des stratégies d'investissements sur de portefeuilles d'actions qui surperforment le rendement normal du marché en utilisant une variété de modèles. Pour cela, les investisseurs collectent les données produites par les différents systèmes financiers pour en trouver des liens et des structures de comportement aidant à la prédiction et à la sélection des meilleurs titres. Or, la nature complexe de données, la vitesse de production des données, le volume et la grande dimension constituent une entrave à l'application des modèles classiques aux données pour en extraire de la connaissance. De ce fait, d'autres modèles de « Data-mining » prennent de la place dans le domaine de la prédiction et de la sélection des titres. Cet article va présenter une revue de littérature sur la question de la prévisibilité de la performance des actions ainsi qu'un aperçu sur les modèles classiques et les modèles de « data-mining » les plus utilisés dans l'évaluation et la prédiction de la rentabilité des actions.

Mots clés : Science de données, séries temporelles, prédiction des titres, fouille de données, apprentissage automatique.

Abstract:

Many researches has been done on the issue of predictability and performance of stocks. Investors have long sought to find investment strategies in stock portfolios that outperform normal market performance using a variety of models. To this end, investors collect the data produced by the various financial systems to find patterns helping to predict and select the best securities. However, the complex nature, the speed of production, the volume and the large dimension of data constituting an obstacle to the application of classical models to data in order to extract knowledge from it. Hence, other "data mining" models are taking up space in the area of prediction and stock selection. This article will present a literature review on the question of the stock performance predictability as well as an overview on the classical and "data-mining" models most used in the evaluation and the prediction of stock profitability.

Keywords: Data-science, time series, stock prediction, data-mining, machine-learning.

Introduction

Nombreux sont les titres offerts au marché financier à travers plusieurs plateformes de « trading », les investisseurs devraient choisir parmi ces titres pour constituer leurs portefeuilles. L'abondance des titres, et l'abondance de l'information qui les entoure, rendent la tâche difficile aux investisseurs à choisir les bons titres. Ce n'est pas tous les titres dans le marché en globalité, ou dans un secteur en particulier, génèrent des profits, c'est la raison pour laquelle les gestionnaires de portefeuilles, dans leur processus de décision d'investissement, utilisent divers modèles pour pouvoir sélectionner les titres afin de construire leurs portefeuilles et ce, dépendamment de la stratégie d'investissement adoptée. Par ces modèles, les investisseurs cherchent à résumer l'information financière et trouver les critères les plus pertinents de comparaison entre les titres pour en choisir les meilleurs.

Le souci de tous les investisseurs est de construire des stratégies d'investissement sur des portefeuilles de titres soigneusement choisis en vue de générer une performance supérieure au marché. La question de mener une gestion active ou une gestion passive a suscité beaucoup de débats depuis longtemps jusqu'à présent. Les militants de la gestion passive croient que la gestion active n'apporte aucune valeur ajoutée et ne surperforme pas plus qu'une simple stratégie suiveuse d'un indice boursier représentant le marché. Cependant, les gestionnaires actifs croient que l'utilisation de leurs talents dans le processus de sélectivité des titres et l'utilisation des modèles de prédiction, apporte de la performance excédentaire par rapport à l'indice de référence.

Depuis l'introduction des modèles mathématiques en finance qui remonte au début du 20^e siècle, ces modèles ont été utilisés tant pour expliquer la performance des titres que pour sélectionner les bons titres en se basant sur leur historique de prix et sur leurs comportements passés. Toutefois, l'utilisation de l'historique de prix pour constituer des stratégies d'investissements surperformant le marché et dégageant un rendement anormal a suscité le débat sur l'efficience des marchés.

Quoi qu'il en soit, la modélisation en finance, et spécialement en gestion de portefeuille, continue son développement en faisant naître de nouvelles spécialités et d'expertises. Du coup, le défi majeur des investisseurs est de trouver une réponse aux principales questions suivantes : 1) quels sont les titres à choisir dans son portefeuille (sélectivité); 2) quels sont les critères sur lesquels se basent la décision d'investissement (règles et données); et 3) à quel

moment l'opération d'achat ou de vente doit être effectuée (timing). Pour répondre à ces questions, les gestionnaires de portefeuilles utilisent une panoplie de modèles, appliqués sur une variété de types de données, à savoir : les modèles qualitatifs, quantitatifs, comportementaux, techniques, fondamentaux, statistiques, économétriques, mathématiques, computationnels, etc.

Par ailleurs, et avec l'essor de technologies de l'informatique et les nouvelles technologies de l'information et de communication (les ordinateurs, les serveurs, les processeurs, les disques de stockage, les algorithmes, l'internet, l'infonuagique, etc.), notre société est devenue un champ vaste de données. Or, le défi n'est plus comment se procurer de la donnée, mais plutôt, comment extraire de l'information utile à partir de ces données. La malédiction de la dimension et la voluminosité de données rendent la tâche difficile aux managers à prendre une décision justifiée en se basant sur leurs modèles classiques. D'où l'apparition de la « science de données » qui regroupe toutes les étapes du processus de traitement de données jusqu'en extraire de la connaissance.

En plus, depuis l'apparition des systèmes de « trading » automatique (algorithmes de trading) dont les règles d'exécution des transactions et la gestion des ordres d'achat et de vente sont implémentés directement dans des programmes, la vitesse de traitement de l'information et la rapidité de la prise de décision deviennent une nécessité et ne sont plus un choix. Ainsi, pour tirer profit de cette évolution technologique et algorithmique, l'introduction d'une approche « machine-learning » dans le processus de prise de décision d'investissement (sélection des titres et les moments d'exécution) regagne en efficacité. Le bénéfice de telle approche automatique permet d'augmenter la précision et de réduire les erreurs causées par les émotions humaines d'une part, et d'exécuter au bon moment les transactions avec leurs meilleurs prix d'autre part.

La science de données est un carrefour entre plusieurs disciplines : les statistiques, les mathématiques, la gestion des bases de données, le « data-mining » (feuille de données) et le « machine learning » (apprentissage automatique). Grâce à ses outils, la science de données est capable de traiter des données volumineuses et de leur appliquer des modèles complexes pour trouver des « patterns » et pour prédire leurs comportements.

Dans cet article, la deuxième partie va définir les concepts les plus utilisés dans le domaine de « data-mining », la troisième partie va retracer l'historique sur la question de la prévisibilité

des prix d'actions dans le cadre d'une gestion active, la quatrième partie proposera un aperçu sur les modèles statistiques et de « machine-learning » les plus utilisés dans la littérature pour l'évaluation et la prédiction de la performance des actions, la cinquième partie mettra l'accent sur les différents types de classifications des modèles « machine-learning » et une seizième et dernière partie traitera les modèles « machine-learning » les plus utilisés en prévision de la performance des actions.

I- DÉFINITIONS DES CONCEPTS

1.1. Intelligence artificielle

L'Intelligence Artificielle (AI) est une science qui vise à utiliser les données pour proposer des solutions aux problèmes existants. C'est la science et l'ingénierie de la reproduction, voire du dépassement du niveau humain, de l'intelligence dans les machines. Le processus de l'AI est enchaîné en six étapes : 1) observer en identifiant les « patterns » dans les données; 2) planifier en trouvant toutes les solutions possibles; 3) optimiser par trouver la meilleure solution; 4) passer à l'action en exécutant la solution optimale; et 5) comparer entre les résultats obtenus et ceux attendus en adaptant la meilleure solution.

1.2. Apprentissage automatique

L'apprentissage automatique (ML) est un sous ensemble de l'Intelligence Artificielle, il est défini comme étant une collection d'algorithmes et des techniques utilisés pour créer des systèmes informatiques capables d'apprendre à partir des données afin de faire des prédictions et des inférences (Swamynathan, 2017).

En 1959, Arthur Samuel, l'un des pionniers de l'intelligence artificielle, a donné la première définition au Machine-Learning (ML) comme étant un champ d'étude visant à donner la capacité à une machine d'apprendre sans être explicitement programmée. Par ailleurs, une définition plus précise du ML a été proposée par Tom Mitchell, de l'université de Carnegie Mellon en 1997 : « A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E ». Avec l'apprentissage automatique, on passe d'une informatique impérative basée sur des hypothèses à une informatique probabiliste basée sur des informations réelles (Biernat & Lutz, 2016).

Les deux principales composantes permettant au ML d'atteindre son objectif sont les deux outils de reconnaissance de comportement, à savoir : les algorithmes d'apprentissage et les modèles d'apprentissage. Avec ces outils, l'apprentissage automatique permet de chercher les « patterns » (comportements et structures) pour comprendre les caractéristiques de données, pour en extraire les connaissances, pour faire des prédictions ou pour identifier les groupes de données (classification). Le ML cherche le modèle qui ajuste le mieux les données avec l'ensemble de réponses correspondantes d'une part, et trouve la façon de comment entraîner et valider le modèle pour qu'il apprenne les caractéristiques du système à partir de données en d'autre part (Suthaharan, 2016).

1.3. Modèles et algorithmes d'apprentissage

Un modèle est un résumé global des relations entre variables, permettant de comprendre des phénomènes, et d'émettre des prévisions. Il se réfère toujours à la modélisation mathématique et statistique. Le modèle cherche à trouver les relations paramétrées entre les données d'une part et l'ensemble des réponses d'autre part (la variable d'intérêt, la cible, la connaissance, les classes, etc.). Ces relations peuvent être des fonctions ou des processus paramétrés qui lient les caractéristiques du système à partir des données d'entrée étiquetées (labellisées). Tandis que les algorithmes d'apprentissage, dans le contexte de l'apprentissage automatique, sont utilisés pour entraîner, valider et tester le « modèle » en utilisant un échantillon de données, et ce, pour trouver les valeurs optimales de ses paramètres, de le valider et d'évaluer sa performance (Suthaharan, 2016).

1.4. Data-mining (fouille de données)

Le terme « Data-mining », qui signifie la fouille de données ou l'exploration de données, a été introduit dans les années 1990 dans la communauté des bases de données. L'évolution des techniques d'exploration de données a commencé parallèlement avec le début de stockage de données sur ordinateurs dans les bases de données relationnelles.

Fayyad et Piatetski-Shapiro (1996) ont défini le DM comme suit: “le « data-mining » est le processus non trivial d'identification, à partir des données, des structures et comportements valides, nouveaux, potentiellement utiles et finalement compréhensible” (traduction libre de Fayyad, Piatetsky-Shapiro & Smyth, 1996). Le DM est une étape dans le processus d'extraction de connaissance « Knowledge Discovery in Databases » (KDD). Bien que le terme «

Knowledge Discovery in Databases » a été inventé par Piatetsky-Shapiro en 1989, il n'a été présenté qu'en 1996 dans le livre de Fayyad et al. Ce dernier a défini le processus KDD comme un processus qui intègre de multiples technologies dans la gestion de données, à savoir : l'entrepôt de données (Data-warehousing), les statistiques, l'apprentissage automatique, l'aide à la décision, la visualisation et le traitement parallèle (Fayyad, et al., 1996).

Le processus d'extraction de connaissances KDD comprend cinq étapes : 1) la sélection de données; 2) le prétraitement des données en collectant et nettoyant les données requises; 3) la transformation de données dans des formes appropriées pour l'exploration : cette étape peut comprendre le lissage, l'agrégation, la normalisation et la réduction de dimension de données; 4) l'exploration de données en appliquant des algorithmes aux données pour en extraire les liens et les relations utiles (patterns), plusieurs méthodes peuvent être utilisées dans cette étape comme l'utilisation des statistiques descriptives pour résumer les données, les techniques graphiques pour visualiser les données, le « Clustering » ou les modèles prédictifs comme la régression et la classification; et 5) l'interprétation et l'évaluation des patterns explorés pour la rendre compréhensible par l'utilisateur (Swamynathan, 2017).

1.5. Analyse de données (Data-analytics)

L'analyse de données ou l'analyse d'affaires (« Data Analytics » ou « Business Analytics ») a été développée vers la fin des années 1960 avec l'introduction des ordinateurs dans l'aide à la prise de décision au sein des organisations. En fait, avec l'évolution des entrepôts de données (Data-warehousing) et le système de Planification de Ressources d'Entreprise (ERP), les dirigeants d'entreprises examinent leurs hypothèses et leurs stratégies d'affaires en analysant les données historiques. L'analyse de données a été bien évoluée avec l'utilisation de l'intelligence d'affaire (Business Intelligence) dans le processus d'aide à la décision.

L'analyse de données est classée en quatre types d'analyses : 1) l'analyse descriptive qui décrit le passé en décrivant et résumant les données dans un format facile à interpréter par l'homme; 2) l'analyse diagnostique qui examine les données par des outils informatiques pour comprendre ou résoudre certains problèmes; 3) l'analyse prédictive qui estime la vraisemblance des événements futurs en se basant sur des techniques comme le « data-mining », les statistiques, la modélisation, « machine-learning » et l'intelligence artificielle; 4) l'analyse prospective qui analyse et optimise l'impact des différents scénarios prédits sur les résultats en

utilisant les règles d'affaires avec des outils comme le « machine-learning », la programmation linéaire, la programmation dynamique, etc. (Swamynathan, 2017).

1.6. Science de données (Data-Science)

En 1960, Peter Naur a utilisé le terme « science de données » dans sa publication « Concise Survey of Computer Methods », qui traite des méthodes contemporaines de traitement de données dans des larges gammes d'applications. Bien que le terme "science des données" a vu le jour depuis l'année 1960, il n'est devenu populaire qu'en 2008 quand Jeff Hammerbacher et DJ Patil, de Facebook et LinkedIn, ont choisi le mot « data-scientist » en décrivant leurs équipes et leur travail (Patil, 2011; Swamynathan, 2017).

Éric Biernat et Michel Lutz (2016) ont défini la « data-science » comme étant une démarche empirique qui se base sur des données pour apporter une réponse à des problèmes (Biernat & Lutz, 2016). De même, S. Suthaharan (2016) a défini la science de données comme étant « la gestion et l'analyse des données, l'extraction de l'information utile, et la compréhension du système produisant la donnée. Ce dernier peut être une seule unité (par exemple un réseau) constituée de plusieurs sous-unités interconnectées (ordinateurs ou capteurs) collaborant sous un ensemble de principes et stratégies pour effectuer des tâches comme la collecte de données, des faits, ou des statistiques de l'environnement que le système est censé contrôler » (traduction libre de Suthaharan, 2016)

Selon le même auteur, l'exécution des tâches de collecte de données, des faits ou des statistiques du système passe par la transformation de la donnée en connaissance (Fig.1). Cette dernière peut être décrite comme une information apprise acquise à partir de données. La connaissance est appelée aussi la réponse du système, elle est obtenue en appliquant un modèle « f » aux données, elle pourrait être la détection des « patterns » dans les données, le calcul d'une distribution inconnue, classification d'une variété de « patterns », calcul de corrélation dans les données, etc. En plus de ça, le contrôle et les transformations peuvent être de nature physique, mathématique ou logique.

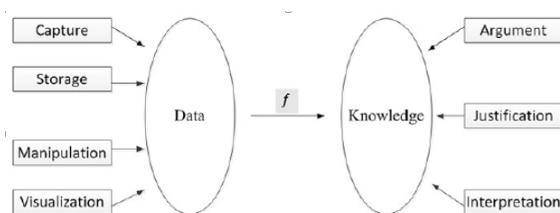


Fig. 1 : Transformation de données en connaissances. (Source : Suthaharan, 2016)

Le cycle de vie d'un projet « data-science », tel qu'il est développé par Venkat Ankam (2016), est composé des huit étapes : 1) identifier le problème d'affaire à résoudre ainsi que les résultats à atteindre; 2) identifier le besoin en terme de données (qualité, quantité, format, sources, etc.); 3) collecter les données; 4) prétraiter les données par la manipulation, l'organisation, la conversion, le nettoyage et l'imputation des données manquantes; 5) poser les hypothèses et procéder à la modélisation : dans cette étape le « data-scientist » formule les hypothèses possibles en fonction du problème à partir desquelles il détermine le modèle approprié et la nature des données à utiliser; 6) mesurer l'efficacité en exécutant le modèle choisi sur les données et en vérifiant les résultats obtenus par rapport aux résultats escomptés : dans cette étape le « data-scientist » définit les métriques de mesure de l'erreur/précision du modèle (Mean Squared Error, MSE) en entraînant le modèle sur des données test; 7) apporter les améliorations nécessaires en fonction des mesures dégagées; et finalement, 8) communiquer les résultats sous forme des rapports et des tableaux de bord (Ankam, 2016).

II- BASE THÉORIQUE SUR LA PRÉVISIBILITÉ DE LA PERFORMANCE DES ACTIONS ET LA GESTION ACTIVE

Depuis le début des années 60, la question sur le processus qui détermine l'évolution des cours des actifs boursiers a été mise en avant. Plusieurs auteurs se posent la question sur l'indépendance des changements successifs dans les prix d'actions et sur leur dynamique de prix si elle évolue en suivant une marche aléatoire (Random Walk). Selon cette dernière, la connaissance de l'historique des prix d'actions n'apporte aucune information additionnelle à la prédiction des rendements anormalement élevés (Rendement excédentaire après rémunération de facteurs de risque), et le travail d'un « chartiste » qui se base sur les cours historiques des titres et de l'analyse technique n'apporte aucune valeur.

Selon la théorie de l'efficience des marchés (prolongement de la théorie du marché pur et parfait des néo-classiques du XIX^e siècle) développée par Fama (1970), il n'est pas possible de développer une stratégie d'investissement capable de « battre le marché » en réalisant un rendement supérieur à celui-ci, car les prix sont transparents et incorporent toute l'information relative aux titres. Selon cette théorie, les prix des actifs financiers reflètent toute l'information disponible, ce qui implique que les changements successifs des prix sont indépendants. À partir de là, il est impossible de réaliser un rendement excédentaire élevé, car les actions s'échangent à leur juste valeur et il n'y a aucune existence des actions surévaluées ou sous-évaluées (Fama, 1970).

Selon Fama (1970) et Roberts (1967), l'efficience peut prendre trois formes : 1) une forme faible qui stipule que l'historique des rendements des titres est incorporée dans son rendement actuel, à cet effet, l'utilisation de l'analyse technique n'apporte aucune prévisibilité sur les rendements des titres; 2) une forme semi-forte dont les prix des titres incorporent toute l'information publique disponible, ce qui implique que les modèles basés sur l'analyse fondamentale fondée sur les informations publiées ne permettent pas d'améliorer la prévisibilité des rendements de titres; et 3) une forme forte selon laquelle les prix reflètent toute l'information publique et privée, de ce fait, aucune information, même si privilégiée, ne peut prévoir les cours futurs des titres (Fama, 1970; Roberts, 1967).

Cependant, la théorie de l'efficience des marchés est très normative, car elle se base sur le fait que les investisseurs ont accès à la même information, ils l'utilisent de la même manière, ils ont les mêmes opportunités d'emprunt et de placement et ils prennent toujours leurs décisions sur une base rationnelle en maximisant le profit tout en minimisant le risque. Or, plusieurs études empiriques viennent démontrer qu'il est possible de constituer des stratégies d'investissements actives permettant de battre systématiquement le marché (Jegadeesh, 1990; O'Shaughnessy, 1997; Rouwenhorst, 1998; Swinkels, 2004). Dès lors, plusieurs études se sont penchées d'expliquer ces anomalies du marché en cherchant les facteurs de risque influençant les rendements, dits anormaux, des titres boursiers, à savoir : le facteur taille des compagnies, le facteur marché, le facteur croissance, le facteur valeur, etc. Fama et French (1996) ont justifié les rendements anormaux des titres en utilisant quatre facteurs de risque : 1) le ratio cours/bénéfice; 2) le ratio cours/valeur comptable; 3) la capitalisation de la compagnie; et 4) le beta du marché (Fama & French, 1996).

Plusieurs auteurs par la suite viennent remettre en cause la théorie d'efficience du marché en analysant l'effet Momentum des titres (Jegadeesh & Titman, 1993). Ils démontrent que les titres qui ont une forte performance dans le passé (dernière année) continuent à mieux performer dans le futur (les six prochains mois) que les titres à faible performance passée, communément appelé « la résistance à la performance » (Carhart, 1997; Sharpe, 1964).

C'est Lévy (1967) qui a mis l'emphasis sur le principe de Momentum ou de la force relative du prix (Levy's trading rules). Il montre que les mouvements de prix ne se font pas selon une marche aléatoire qui stipule l'imprévisibilité totale des prix futurs. Par son résultat, largement controversé, les actions présentant un rendement supérieur à la moyenne du marché continuent à offrir le même rendement dans le futur (Levy, 1967).

Dès lors, plusieurs études viennent appuyer l'effet du Momentum. Jegadeesh (1990) montre que la stratégie qui se base sur la sélection des titres en fonction de son rendement des 12 derniers mois surperforme l'indice du marché de référence (Jegadeesh, 1990). En outre, O'Shaughnessy (1997) a vérifié que la stratégie basée sur le Momentum des prix permet de battre le portefeuille de référence. En effet, un portefeuille de 50 titres sélectionnés sur la base du Momentum obtient un rendement supérieur de 1.6% par rapport à l'indice de référence S&P500 (O'Shaughnessy, 1997).

Dans le même cadre des critiques de la théorie d'efficience des marchés, Grossman (1976) et Grossman et Stiglitz (1980) ont argumenté l'impossibilité d'existence des marchés parfaitement efficients par le besoin des agents de collecter l'information. D'après ces auteurs, les marchés doivent avoir certains degrés d'inefficience pour compenser les investisseurs pour leur collecte d'information, sinon, il n'y aura aucun gain de la collecter (Grossman, 1976; Grossman & Stiglitz, 1980). Lo (2004) a rejeté l'hypothèse de la marche aléatoire de la dynamique de prix des actions. Il a conclu la dépendance des rendements actuels avec les rendements passés des actions, ce qui laisse comprendre que la prédiction des rendements futurs est possible en se basant sur leurs valeurs historiques (Lo, 2004).

La finance comportementale (behavior finance) montre que les décisions d'investissements prises par les agents économiques ne sont pas rationnelles en leur totalité. Elles partent de l'hypothèse qui stipule que si tous les investisseurs sont rationnels, ils vont vendre les mêmes titres en même temps, ce qui contredit à la notion du marché lui-même, qui suppose la présence des acheteurs et des vendeurs. Ces études montrent que les décisions d'investissement sont

influencées par le comportement humain, à savoir : les émotions, le mimétisme (imiter les stratégies des autres), le conservatisme (réaction trop lente aux nouvelles informations), l'excès de confiance, etc.

Ces anomalies de marché laissent la place à la gestion active des arbitragistes. Ces derniers exploitent ces irrégularités (opportunités d'arbitrage) pour tirer profit du marché. Ils misent sur l'existence des rendements anormaux sur le marché en développant des différentes stratégies. Ils agissent comme étant des acteurs qui contribuent au retour à l'équilibre des marchés.

III- MODÈLES D'ÉVALUATION ET DE PRÉDICTION DE LA PERFORMANCE DES ACTIONS

La littérature est très riche des modèles de la sélection et de la prédiction des titres, Akhter Rather et al. (2017) ont pu classer plus de 140 recherches les plus importantes dans le domaine de la prédiction des titres et de la sélection de portefeuilles depuis l'année 1926 jusqu'à l'année 2015. Ils ont classifié par ordre chronologique les articles selon s'ils traitent un modèle de prédiction ou de sélection de portefeuilles, et si c'est un modèle unique ou hybride (Rather, Sastry & Agarwal, 2017).

Notamment, avant de prendre une décision d'investissement dans les actifs financiers et pour réduire l'incertitude autour des prix futurs des titres afin d'avoir une visibilité sur leurs investissements, les actionnaires, les investisseurs et les gestionnaires de portefeuilles ont besoin de s'informer sur la qualité de ces titres en essayant d'en expliquer la formation du rendement. Ils utilisent pour cette raison l'information financière qu'ils disposent sur l'évolution des prix des titres d'une part, et l'information relative aux firmes émettrices de ces titres d'autre part, et ce, selon deux types d'approches d'analyse : l'analyse fondamentale et l'analyse technique.

3.1. Modèles basés sur l'analyse fondamentale

Les analystes fondamentaux examinent les faits qui affectent la valeur et la croissance des firmes, ils analysent les états financiers de la firme, les ratios financiers, les opinions des experts, les décisions du conseil d'administration, le prix de l'or, la croissance économique, le taux d'inflation, le taux d'intérêt, la production industrielle, etc. Ils utilisent ces variables parce

qu'elles fournissent de l'information sur la valeur des actions de l'entreprise par rapport à sa croissance potentielle des bénéfices futurs. Elles peuvent être également considérées comme des indicateurs sur la réalisation du potentiel de la croissance des actions.

Les rapports annuels ou semestriels (états financiers et autres) sont riches d'informations financières que les analystes, chercheurs et investisseurs peuvent l'exploiter en la transformant en ratios financiers. Ces derniers permettent aux investisseurs de former une base d'information sur la firme dans leur prise de décision d'investissement. Par ailleurs, l'analyse des ratios a été le paramètre clé utilisé par les gestionnaires des fonds et les investisseurs dans la valorisation des actions pour trouver leurs valeurs intrinsèques. Aujourd'hui, les ratios financiers sont largement utilisés dans l'analyse fondamentale pour prédire la performance future des firmes (Dutta, Bandopadhyay & Sengupta, 2012).

3.2. Modèles basés sur l'analyse technique

Les analystes techniques utilisent les modèles qui examinent le passé pour anticiper l'évolution future des prix en suivant l'évolution du marché. Ils utilisent des variables dérivées des séries temporelles des prix comme : l'indice de force relative (Relative Strength Index, RSI), le Momentum ou oscillateur stochastique, le volume des transactions, le prix de clôture, prix d'ouverture, prix haut, prix bas, les valeurs historiques de prix, etc. L'idée principale derrière l'utilisation de ces indicateurs est d'évaluer les mouvements des cours des actions en fonction des tendances et des volumes historiques des prix.

3.3. Modèles stochastiques

3.3.1. Marche aléatoire

La modélisation des prix d'actions sous forme d'une marche aléatoire est le résultat direct de la théorie de l'efficience des marchés. La marche aléatoire est définie comme étant la somme des variables aléatoires indépendantes et identiquement distribuées (i.i.d). Roberts (1959) a trouvé qu'une série temporelle des prix d'actions ressemble à une série numérique composée d'un cumul de nombres aléatoires, et que la différence première des nombres générés aléatoirement ressemble aussi à la différence première des cours boursiers. Les marches aléatoires sont des processus markoviens sans mémoire, c'est-à-dire, la valeur dans le futur étant donné le passé et le présent ne dépend que du présent (Roberts, 1959).

3.3.2. Mouvement brownien (ou processus de Wiener)

Le mouvement brownien réfère aux travaux de Robert Brown (1827) qui tente de modéliser le mouvement des particules dans un liquide, et c'est le mathématicien Norbert Wiener qui a analysé rigoureusement les mathématiques derrière ce processus stochastique (Taylor & Karlin, 2014). Plus tard, Osborne (1959) a introduit ce processus brownien de la physique à la finance. À l'instar de Robert (1959), Osborne constate que les logarithmes des changements de prix sont indépendants les uns aux autres. Il établit une ressemblance entre la dynamique des prix d'actions et le mouvement brownien (Osborne, 1959).

Black et Scholes (1973) ont exprimé la dynamique des prix des actions sous forme d'un mouvement brownien géométrique suivant une loi log-normale. En suivant cette dynamique des prix, les rendements des actions sont des mouvements browniens arithmétiques suivant une loi normale. À partir de cette modélisation des prix des actions découle beaucoup de modèles d'évaluation d'options sur les actions (Black & Scholes, 1973; Merton, 1974).

Toutefois, l'hypothèse de l'indépendance et la normalité des rendements du modèle Black et Scholes ont suscité beaucoup de critiques (Genest, Ghoudi & Rémillard, 2007).

3.3.3. Volatilité stochastique

Pour pallier le problème de la volatilité constante rencontré dans le modèle Black et Scholes (1973), les modèles à volatilité stochastique ont vu le jour. Ils ont été introduits par Engel (1982) par son modèle ARCH et ils ont été généralisés par des modèles GARCH de Bollerslev (1986), inspirés de la démarche de Box et Jenkins (1970) avec une dynamique autorégressive. En effet, la famille des modèles ARCH permet d'avoir des processus avec des queues de distribution plus épaisses (distribution leptokurtique) comparativement au processus ARMA (Bollerslev, 1986; Engle, 1982).

3.4. Modèles statistiques simple

Avant de sélectionner les titres et de choisir la meilleure stratégie d'investissement, les investisseurs ont recours à des modèles statistiques simples pour comparer la profitabilité des titres et pour expliquer la formation du rendement des actions en introduisant plusieurs facteurs de risques, à savoir : facteurs macroéconomiques, facteurs fondamentaux, facteurs statistiques et facteurs techniques.

Autrement dit, ces modèles cherchent à trouver les facteurs de risque expliquant le mieux le rendement d'actions moyennant des méthodes simples de régression linéaire ou d'optimisation. Parmi les modèles les plus utilisés dans la littérature, on trouve:

- Modèle à un facteur : Le modèle d'évaluation d'actif financiers MEDAF (CAPM) met une structure sur la théorie de l'approche moyenne-variance de Markovitz (Markowitz, 1952). Il présume que le risque systématique du marché est le seul risque expliquant le rendement.
- Modèle à trois facteurs de Fama et French (1993) : C'est un modèle qui utilise les facteurs expliquant le rendement, à savoir : le rendement excédentaire de marché, la prime SMB (Small, Medium, Big) qui représente le rendement excédentaire des titres de petite capitalisation par rapport aux titres de grande taille, et la prime HML (High, Medium, Low) qui reflète le rendement excédentaire des actions de valeurs par rapport aux actions de croissance (Fama & French, 1993).
- Modèle à quatre facteurs de Carhart (1997) : C'est le modèle à trois facteurs de Fama et French plus le facteur Momentum (Carhart, 1997).

3.5. Modèles économétriques de séries temporelles

Les modèles économétriques cherchent les relations linéaires (corrélation et autocorrélation) entre les valeurs historiques des prix d'une série chronologique et ses valeurs présentes pour prédire le futur. La prédiction de la rentabilité des titres par les modèles économétriques en se basant sur leur historique suppose la stationnarité des rendements des titres. Néanmoins, la stationnarité à court terme a été justifiée par le travail de Starica & Granger (2005), qui ont évoqué la notion de la stationnarité locale des actions, et le travail de Jegadeesh & Titman (1993) qui ont relevé l'effet Momentum dans le prix d'actions.

Les modèles économétriques font partie des modèles statistiques, ils se sont utilisés principalement en prédiction de la dynamique des séries temporelles. À titre d'exemple, les modèles de lissage exponentiel sont largement utilisés grâce à leur valeur apportée en prédiction dans la pratique, ils prédisent la valeur d'une variable aléatoire en pondérant décroissement ses valeurs passées dans la série temporelle.

En outre, le processus ARMA développé par Box et Jenkins (1970) est considéré l'un des modèles les plus répandus dans le domaine de la prédiction des séries temporelles, il gagne plus de place tant sur le plan théorique que sur le plan pratique. Sa capacité à décrire l'évolution

des séries temporelles le rend comme étant l'outil puissant pour la prévision des séries chronologiques, il prédit le mouvement futur d'une série temporelle en utilisant ses observations passées (Box & Jenkins, 1970).

En effet, le modèle ARMA est la combinaison du processus AR (Autoregressive) développé par Yule (1926), et du processus MA (Mobile Average) développé par Wold (1938) (N. & Wold, 1939; Yule, 1926). Comme son nom l'indique, le processus AR est une régression de la valeur actuelle contre les valeurs antérieures de la variable elle-même dont le nombre de valeurs retardées incluses dans le modèle est défini par son degré « p ». Selon le processus AR, la valeur future peut être expliquée par les observations passées. Tandis que le processus MA explique la valeur future par les erreurs aléatoires passées. ARIMA est une version plus avancée du modèle ARMA, car il intègre l'étape de différenciation première de la série (Integrated) pour la rendre stationnaire.

Selon la méthodologie Box et Jenkins (1976), le processus de modélisation ARIMA (p, d, q) passe par quatre étapes : 1) l'étape d'identification qui consiste à trouver l'ordre de la composante AR, « p », et l'ordre de la composante MA, « q », en utilisant le diagramme d'autocorrélation (ACF), le diagramme d'autocorrélation partielle (PACF) et les critères d'information AIC et SBIC. La détermination du nombre de différenciations premières « d » pour « stationnariser » la série est généralement déterminée par le test Dickey-Fuller Augmenté (ADF) qui teste l'existence des racines unitaires dans la série. La série qui dispose une racine unitaire doit être différenciée autant de fois jusqu'à ce qu'elle soit stationnaire en n'ayant aucune racine unitaire; 2) l'étape d'estimation des paramètres : dans cette étape, la méthode de moindre carré ordinaire (Least Square) ou la méthode du maximum de vraisemblance (Maximum Likelihood) sont utilisées pour l'estimation des paramètres; 3) l'étape de validation du modèle; pendant laquelle, on vérifie si le modèle est adéquat en diagnostiquant les résidus; et finalement, 4) l'étape de la prévision (Box & Jenkins, 1970).

Toutefois, pour analyser les effets mutuels et les interdépendances entre les variables à un moment donné, et si les variables ont un caractère autorégressif (c.à.d. les données passées d'une variable influencent son état actuel), les modèles univariés comme AR, MA et ARIMA ne captent pas ces interactions, alors que les modèles VAR (vecteur autorégressif) viennent résoudre ce problème. Le modèle VAR explique les valeurs actuelles de plusieurs variables de telle sorte que la valeur actuelle de chaque variable est expliquée par les valeurs

actuelles des autres variables d'une part, et les valeurs retardées des variables elles-mêmes et les autres variables d'autre part.

3.6. Modèles « Data-mining »

Le volume, la vitesse et la variété (3V) des données produites par les marchés et tous les systèmes financiers ont attiré les chercheurs à explorer des nouvelles méthodes pour développer des stratégies d'investissements performantes. De ce fait, les « data-scientist » adoptent des approches « data-mining » (DM) par collecter les données, les nettoyer, les réduire, les explorer et les traiter pour chercher les liens, les dépendances et les corrélations entre les différentes variables en vue d'en extraire les connaissances utiles à la prise de décision d'investissement.

Le « machine learning » (ML) est un outil utilisé, dans le cadre de DM, par les « data-scientist » pour chercher les « patterns » et les relations systématiques entre les variables. Pour extraire de la connaissance à partir des données, le DM utilise la reconnaissance des patterns, la classification, le Clustering, les associations, les probabilités, l'optimisation, et bien d'autres méthodes.

La première application des techniques de ML dans la prévision de la performance des titres date depuis 1988, quand les réseaux de neurones artificiels (NNA) ont été appliqués à la prévision des actions. Un réseau de neurones sans rétroaction (Feed Forward Network) a été utilisé pour analyser les rendements quotidiens des actions d'IBM. Dès lors, beaucoup de recherches viennent appliquer les techniques de ML pour tenter de trouver des règles prédictives du marché d'actions et bien d'autres multitudes d'actifs financiers (Booth, Gerding & McGroarty, 2014).

Keerti. S. Mahajan et al. (2013) ont mis le point sur les applications de différentes techniques de DM au marché financier d'actions (Mahajan & Kulkarni, 2013). La machine à vecteurs du support, les arbres de décisions, les règles d'association, le plus proche voisin, les forêts aléatoires, les réseaux de neurones et bien d'autres méthodes ont été utilisés pour la sélection, l'évaluation et la prédiction des titres.

3.7. Modèles « Data mining » versus modèles statistiques

Le DM ne traite pas d'estimation et de tests des modèles préspecifiés comme les statistiques, mais il traite la découverte de modèles à l'aide d'un processus de recherche algorithmique

d'exploration de modèles (linéaire ou non-linéaire, explicites ou implicite). Les modèles dans le DM ne sont pas issus d'une théorie, mais de l'exploration de données. Les deux types de techniques, statistiques et d'apprentissage automatique, ont le même objectif qu'est l'apprentissage à partir de données. Néanmoins, les techniques de l'apprentissage automatique ne sont pas guidées par la théorie économique et statistique. Les méthodes statistiques traditionnelles se sont basées sur des hypothèses statistiques (par exemple la normalité des distributions), tandis que les méthodes ML se concentrent surtout sur l'obtention des meilleures prédictions.

La finance a largement utilisé les méthodes statistiques traditionnelles comme la régression linéaire qui ajuste une ligne droite aux données. Or, la nature de données est généralement non linéaire. Nombreuses sont les techniques de ML qui sont capables d'inférer les relations non linéaires à partir des données. Les méthodes de ML se penchent souvent à performer les résultats de la prédiction sur des problèmes de grande dimension contrairement aux méthodes statistiques traditionnelles qui se concentrent sur l'inférence statistique relative aux problèmes de la faible dimension (intervalle de confiance, tests d'hypothèses, meilleur estimateur, etc.) (Wall, 2018).

IV- MACHINE LEARNING : CLASSIFICATION

Les techniques de « machine-learning » peuvent être classifiées sous les critères suivants:

4.1. Apprentissage supervisé, non supervisé et par renforcement

Les méthodes de l'apprentissage automatique sont classées en trois principaux types : apprentissage supervisé, apprentissage non supervisé et apprentissage par renforcement.

Les techniques d'apprentissage supervisé, dites prédictives, sont capables d'apprendre les liens et les relations (patterns) à partir d'un échantillon exemple et avec des sorties exemples (données étiquetées). Les données sont sous forme de couples entrée-sortie dont les sorties (classes ou événements) sont préalablement connues. L'objectif des algorithmes d'apprentissage supervisé est d'apprendre les « patterns » à partir de données en vue de construire un ensemble de règles associant les entrées à des classes (ou à des événements) (Swamynathan, 2017). Elles ajustent les données sur leurs étiquettes (labels) en minimisant

une fonction d'erreur (Erreur quadratique moyenne) et/ou en maximisant la précision. Généralement, les techniques d'apprentissage sont utilisées quand la fonction qui relie les entrées et les sorties est inconnue, et nous disposons que des échantillons exemples avec des entrées et des sorties.

Les techniques d'apprentissage non supervisé, dites exploratoires, sont capables de chercher la structure ou la distribution de données sans être dirigée par des sorties exemples et sans être assistée par l'être humain. Dans ce type d'apprentissage, les sorties sont inconnues pour les données historiques, toutes les données sont équivalentes. L'objectif est de classer les individus dans des groupes qui ont des caractéristiques similaires. Ces techniques traitent les problèmes de Clustering dont les données sont groupées dans des groupes d'observations qui s'appellent « Clusters ».

L'apprentissage par renforcement a pour objectif d'apprendre un comportement face à diverses situations. Il correspond à un agent intelligent qui apprend au fil des expériences en associant des situations à des actions qui offrent le maximum de « récompense » finale (Biernat & Lutz, 2016).

4.2. Classification ou régression

Les techniques d'apprentissage automatique supervisé traitent les problèmes de classification et les problèmes de régression. Quand les données sont étiquetées avec des variables réponses (l'ensemble des connaissances) sous forme de classes dénombrables et d'un nombre bien défini à l'avance (variable à expliquer est de forme catégorielle), on parle de la classification. Mais quand la variable réponse est numérique continue et non pas sous forme de classes, on parle de la régression.

4.3. Mathématique, hiérarchique ou de couches

Les modèles mathématiques utilisent tout le domaine de données pour trouver le bon classificateur. Ils supposent que tout le domaine de données soit défini préalablement et soit disponible pour le développement du modèle et de l'algorithme d'apprentissage, comme par exemple la « machine à vecteurs de support » et la « régression logistique ». Par ailleurs, Les modèles hiérarchiques divisent le domaine et les classes en petits morceaux horizontalement et verticalement et les regroupe plus tard pour former les classes finales. Les modèles hiérarchiques sont similaires aux modèles mathématiques en termes d'utilisation intégrale de

domaine pour trouver le classificateur, comme par exemple les « arbres de décision » et les « forêts aléatoires ». Finalement, les modèles en couches sont différents des autres modèles quant à l'utilisation du domaine de données, ils peuvent utiliser les observations de données séparément, comme ils peuvent fonctionner en apprentissage en ligne. Ils supposent la définition d'une probabilité à l'aide de certaines fonctions mathématiques et ils assignent une probabilité à chaque observation, puis ils l'affectent à une classe comme les « réseaux de neurones » (Suthaharan, 2016).

V- APPRENTISSAGE SUPERVISÉ

5.1. Fonctionnement

Après avoir choisi l'algorithme de la technique ML à appliquer (régression, SVM, arbre de décision ou autre) pour traiter le phénomène à étudier (la réponse au problème), l'échantillon de données disponibles à l'étude doit être scindé en deux principales parties, un jeu de données d'« entraînement-validation » généralement fixé à 80% des données disponibles, et un jeu de données « test » pour le restant des données (20%).

Pour augmenter leur capacité de généralisation, les techniques d'apprentissage automatique fonctionnent selon les étapes suivantes : 1) entraînement; 2) validation; et 3) test. Lors de la première étape d'entraînement, l'algorithme estime, approxime et optimise les paramètres du modèle en minimisant l'erreur entre les valeurs prédites par le modèle et les valeurs réelles des labels (étiquettes). Dans cette étape, l'algorithme utilise le jeu de données d'entraînement et des mesures quantitatives d'évaluation d'erreur comme l'entropie dans le cas de la classification et l'erreur quadratique moyenne dans le cas de la régression. Dans l'étape test, l'algorithme teste la fiabilité du modèle optimisé lors de la première étape d'entraînement en utilisant le jeu de données « test », données étiquetées jamais vues par le modèle. C'est dans cette étape que l'algorithme calcule des mesures qualitatives de performance, comme par exemple la précision, la sensibilité ou la spécificité, pour accepter ou rejeter le modèle selon les résultats obtenus en matière de performance du modèle et de précision de la classification.

L'étape de la validation est une étape importante, elle est intégrée conjointement avec l'étape d'entraînement. Elle vise à tester le modèle entraîné avant qu'il soit testé réellement dans l'étape « test ». Elle est similaire à l'étape « test » quant aux mesures d'erreur calculées, la

seule différence entre l'étape de validation et l'étape « test » est que l'étape de validation est effectuée en plusieurs fois parallèlement avec l'étape d'entraînement, et que les données utilisées dans cette étape sont utilisées alternativement entre entraînement et validation. Pour cette raison, le jeu de données d'entraînement est divisé entre sous-échantillon d'entraînement et sous-échantillon de validation (par exemple, 80% et 20% successivement). L'objectif de la validation est de tester le modèle sous plusieurs conditions pour vérifier la fiabilité des paramètres dérivés lors de la phase d'entraînement. Si le résultat n'est pas satisfaisant, le modèle s'entraîne encore une fois jusqu'à ce qu'il obtienne des meilleurs paramètres. C'est dans cette étape que le modèle corrige très tôt le sur-apprentissage (ou sur-entraînement) qui mène au problème de sur-ajustement (Suthaharan, 2016).

La validation croisée est une technique utilisée lors de la phase de validation du modèle dans le processus d'apprentissage automatique supervisé, elle vise à utiliser plusieurs fois les données entre entraînement et validation pour mesurer d'une manière plus générale la qualité du modèle. Plusieurs méthodes sont utilisées dans la validation croisée (« leave-one-out cross validation », « leave-k-out cross validation », « k-fold cross validation »), la méthode « k-fold cross validation » est une méthode non exhaustive et la plus utilisée, car elle nécessite moins de temps de calcul et approxime les autres méthodes exhaustives. Avec cette méthode, les données sont aléatoirement divisées en k sous-échantillons de tailles égales, parmi lesquels, un sous-échantillon est utilisé pour la prévision et les autres k-1 sous-échantillons sont utilisés pour l'estimation du modèle.

En finance et spécialement dans le cadre de la sélection des titres, de la prévision des prix des actions, de la prévision de la performance de portefeuilles, ou de la prévision de la tendance des prix des titres, les données généralement utilisées sont sous la forme des séries temporelles indexées par le temps. De ce fait, une attention particulière devrait être faite au moment de la validation croisée du modèle. En effet, la validation croisée dans le cas où l'ordre chronologique des observations est important est différente à celle appliquée aux données non temporelles, dans le sens où la prévision doit être appliquée à un sous-échantillon de validation postérieure à celui d'entraînement, donc un choix aléatoire ne fonctionne pas. De ce fait, deux approches sont envisageables : 1) l'entraînement (estimation du modèle) se fait sur un sous-échantillon de données, temporellement ordonnées, dont la taille augmente graduellement, la validation (la prévision du modèle) se fait sur les observations postérieures sur un horizon de prévision donné; et 2) l'estimation se fait sur une fenêtre glissante de taille fixe, alors que la

prévision se fait sur des observations postérieures sur un horizon de prévision donné (Biernat & Lutz, 2016).

5.2. Sur-apprentissage

Dans un monde réel, les données utilisées lors de l'apprentissage comportent du bruit. Alors, au moment de l'apprentissage, les techniques de ML peuvent apprendre plus sur ce bruit et font ce qu'on appelle le « sur-apprentissage », le « sur-ajustement » ou le « surentrainement » (Overfitting), ce qui réduit leur capacité de généralisation sur les nouvelles données.

Un bon modèle est celui qui décrit le mieux les données d'apprentissage et en même temps est capable de générer des meilleures prévisions. Pour cela, les « data-scientist » cherchent le juste équilibre entre la complexité du modèle (minimisation de l'erreur de modélisation) et la capacité de généralisation (minimisation de l'erreur de prévision), dans ce qu'on appelle le compromis biais-variance. Donc, pour éviter le phénomène de sur-apprentissage (ou le sous-apprentissage), une partie de données va être utilisée pour calculer le modèle (jeu d'entraînement) et une autre partie pour effectuer les prévisions (jeu de validation). L'erreur de modélisation diminue avec l'augmentation de la complexité du modèle, tandis que l'erreur de prévision va s'améliorer (diminuer) en augmentant la complexité de modèle jusqu'à un certain niveau d'équilibre à partir duquel elle commence à augmenter également. Quand l'augmentation de la complexité cesse d'améliorer les résultats de la prévision, le modèle commence à surapprendre et perd sa capacité de généralisation (Biernat & Lutz, 2016).

Un sous-apprentissage est caractérisé par une erreur de modélisation élevée, un fort biais (écart entre les données modélisées et les vraies valeurs) et une faible variance, par contre, les prévisions à base d'un modèle sous-apprentis sont stables : des petits changements dans les données ne varient pas trop la prédiction. Cependant, le sur-apprentissage est caractérisé par des modèles à faible erreur de modélisation, faible biais, mais de forte variance avec une erreur de prévision élevée (instabilité en prévision) (Biernat & Lutz, 2016).

VI- MACHINE LEARNING DANS LA SÉLECTION ET LA PRÉVISION DE LA PERFORMANCE DES TITRES

Dans ce qui suit, nous mettrons en avance une revue de littérature sur les techniques les plus utilisées en prédiction de la performance des actions.

6.1. Machine à vecteurs de support (SVM)

SVM est parmi les algorithmes mathématiques de l'apprentissage supervisé les plus puissants et les plus populaires, généralement en « machine learning » (ML) et spécialement en prévision des séries temporelles, car elle s'applique tant aux problèmes de classification qu'aux problèmes de régression. Il y a deux types de SVM : linéaire et non linéaire. La SVM se diffère d'autres méthodes d'apprentissage par le fait qu'elle cherche à construire des modèles fiables en minimisant le risque structurel du modèle (Structural Risk Minimisation), tandis que les autres méthodes cherchent à minimiser le risque empirique (Empirical Risk Minimisation), dite erreur d'entraînement (Rosillo, Giner & De la Fuente, 2014).

Vapnik (1963) a proposé le premier algorithme de la marge maximale des hyperplans dans une SVM linéaire. Par ailleurs, Boser, Guyon et Vapnik (1992) ont trouvé une façon de créer des classificateurs non linéaires en appliquant ce qu'on appelle « l'astuce du noyau », (« kernel-trick », proposé initialement par Aizerman et al. (1964)), pour la maximisation de la marge entre la frontière de décision (l'hyperplan séparateur) et les observations les plus proches appelées « vecteurs de support ». En effet, quand les données sont inséparables linéairement selon les classes de la variable dépendante, l'astuce de noyau permet de les transformer dans un nouvel espace à dimensions supérieures linéairement séparables. En général, quatre groupes de fonctions de noyau peuvent être trouvés : 1) les fonctions linéaires; 2) les fonctions polynomiales; 3) les fonctions radiales; et 4) les fonctions sigmoïdes (Vaiz & Ramaswami, 2016).

Kim (2003) a utilisé la technique SVM pour prédire la direction de la variation du prix quotidien de l'indice composite coréen en la comparant avec les réseaux de neurones retro-propagation (BNN) et le Case Base Reasoning (CBR). Pour son étude, il a employé 12 indicateurs techniques et des données historiques quotidiennes depuis le mois de janvier 1989 jusqu'au mois de décembre 1998. Il a trouvé que le SVM surperforme les autres méthodes en

concluant que la méthode SVM est une technique prometteuse pour la prévision des séries temporelles en finance (Kim, 2003).

Rosillo et al. (2014) ont proposé un système de négociation d'actions basé sur le SVM. Ils ont appliqué un modèle GARCH sur les données quotidiennes de l'indice S&P500 entre l'année 2001 et 2010 pour simuler quatre types de marché : marché à tendance baissière, à tendance haussière, à faible volatilité et à forte volatilité. Leur stratégie hebdomadaire consiste à entraîner le SVM sur des données labélisées en « achat » et « vente » en comparant le prix de clôture de la journée avec les prix des cinq prochains jours qui suivent, et d'autres variables issues de l'analyse technique à savoir, l'indice de force relative RSI et Convergence Divergence Moyenne mobile (MACD). Ils ont trouvé que la stratégie basée sur la SVM performe la stratégie « acheter et garder les actions » (Buy and Hold) et la stratégie naïve, et ce, dans le cas du marché à haute volatilité (Rosillo, et al., 2014).

Sheta et al. (2015) ont comparé la prédiction de l'indice américain S&P500 en utilisant trois méthodes : 1) machine à vecteurs de support (SVM); 2) réseau de neurones artificiels (ANN); et 3) la régression multiple. Leur recherche portait sur une série de 1192 jours de données quotidiennes de l'indice boursier S&P500 entre l'année 2009 et l'année 2014, incluant 27 variables indépendantes de nature technique et fondamentale. Par ailleurs, ils ont intégré aussi dans leur recherche une optimisation sur le nombre de neurones utilisés dans les ANN à perceptron multiple (MLP) à couche cachée unique. Ils ont trouvé le nombre optimal de neurones à utiliser qui est 20 neurones. Ils ont trouvé que le SVM a surperformé les deux autres techniques à savoir, les ANN-MLP et la régression multiple (F., Elsir & Faris, 2015).

Pour prédire le rendement journalier de l'indice BIST100 en Turquie, Oztekin et al. (2016) ont utilisé les réseaux de neurones artificiels (ANN), le système d'inférence fou (fuzzy inference system, FIS) et la machine à vecteurs de support (SVM). Les ANN ont été configurés en algorithme retro-propagation (back-propagation) multicouche à propagation directe (feed-forward) dont l'architecture est (6, n, m, 1) : six neurones en couche d'entrée, deux couches cachées et une couche de sortie. Ils ont trouvé que la SVM superforme les autres modèles (Oztekin, Kizilaslan, Freund & Iseri, 2016).

Kamley et al. (2016) ont examiné 30 articles publiés entre l'année 2000 et l'année 2015 pour identifier les variables fondamentales, techniques et macro-économiques les plus utilisées pour la prévision de la performance des actions d'une part, et les méthodes de prédictions utilisées

pour évaluer la performance des actions en d'autre part. S'agissant des variables fondamentales, douze articles parmi les 30 articles étudiés ont utilisé le ratio cours / bénéfice, le bénéfice par action, la valeur liquidative, l'indice général (IG), le volume des actions, la valeur comptable, la valeur nominale, les nouvelles sur les achats et ventes d'actions, le rendement des dividendes, le taux des bons du Trésor, le ratio courant, le ratio de levier financier, le compte de résultat, la croissance des revenus, la croissance des ventes nettes, la croissance du bénéfice net, le rendement des capitaux propres, la marge bénéficiaire nette, le ratio prix / ventes, etc. Quant aux indicateurs techniques les plus utilisés on trouve : moyenne mobile (MA), moyenne mobile exponentielle (EMA), Indice de force relative (RSI) et Divergence Convergence Moyenne Mobile (MACD), bandes de Bollinger, oscillateur stochastique, Momentum, Slow %D, William R%, taux de variation des prix (ROC), disparité (5-10), oscillateur de prix (OSCP), Commodity Channel Index (CCI), tendance de prix et volume (PVT), etc. Par ailleurs, Ils ont trouvé que les modèles basés sur la SVM se situent au top du classement avec une forte surperformance en précision de prédiction comparativement aux autres modèles à savoir, l'arbre de décision, les réseaux de neurones et la méthode bayésienne (Kamley, Jaloree & Thakur, 2016).

Ouahilal et al. (2017) ont appliqué l'algorithme de SVR (support machine régression) aux données temporelles filtrées (Close, Open, High, Low, Volume), de quelques grandes compagnies marocaines, entre l'année 2004 et 2016, pour prédire le prix des actions. Ils ont utilisé trois filtres pour enlever le bruit à partir des données tout en décomposant la série temporelle en cycles et tendance. Les filtres utilisés sont : le filtre de « Hodrick-Prescott », de « Baxter king » et de « Christiano Fitzgerald ». Ils ont comparé les résultats de la prédiction en combinant le SVR avec chacun des filtres susmentionnés en se basant sur la mesure d'erreur de pourcentage moyenne MAPE (Mean Average Percentage Error). Ils ont trouvé que la combinaison entre le SVR et le filtre de « Hodrick-Prescott » offre le meilleur résultat en termes de réduction d'erreur (Ouahilal, Mohajir, Chahhou & Mohajir, 2017).

6.2. Réseaux de neurones (NN)

La technique des réseaux de neurones (Neural Network) est l'une des techniques prometteuses largement utilisée dans le domaine de la finance. Elle imite le cerveau humain par sa capacité d'apprendre et la capacité de généralisation des problèmes le plus complexes sur de larges données. La technique de réseaux de neurones fonctionne de telle sorte qu'elle affecte des poids

optimaux aux interconnexions entre les variables pour s'ajuster le mieux possible à la variable dépendante en minimisant l'erreur de prédiction. Elle est utilisée à la fois en apprentissage supervisé et non-supervisé. La technique de réseau de neurones est inspirée du fonctionnement des réseaux de neurones biologique, elle traite en parallèle des tâches élémentaires par les différents nœuds pour aboutir à des résultats plus complexes. Sa force réside dans sa capacité à apprendre des expériences passées et de pouvoir détecter les liens entre les variables même s'ils sont complexes (Ehsan, Hamed & Jamal, 2010).

Les réseaux de neurones se sont qualifiés de boîtes noires avec des entrées et une ou plusieurs sorties, à l'intérieur de ces boîtes noires il y a des couches composées de neurones jouant le rôle des calculateurs et des connecteurs entre ces neurones. Un neurone est généralement caractérisé par un vecteur d'entrée donnée par l'utilisateur, un vecteur de poids servant comme des pondérations aux observations d'entrées et qui va être modifié pendant l'apprentissage, une fonction de propagation qui traite de l'information générant un signal d'entrée, une fonction d'activation qui calcule la transformation de l'état précédente vers l'état actuel, et finalement une fonction de sortie qui calcule la valeur de sortie d'un neurone en fonction de son état d'activation.

Il y a des différents types de réseaux de neurones qui se différencient par leurs structures, on trouve : 1) le perceptron multicouche (Multilayer Perceptron, MLP) dont la fonction d'activation est sigmoïdale; 2) les réseaux récurrents; 3) les réseaux modulaires; et 4) les réseaux polynomiaux (Gómez-Ramos & Venegas-Martínez, 2013). Le perceptron multicouche à propagation directe (MLP feed-forward) est le plus populaire, il se différencie des autres méthodes par le fait qu'il établit des connexions avec les neurones postérieurs et interdit la communication avec les couches antérieures (Pas de Feed-Back). Il est capable d'apprendre les fonctions non linéaires très complexes avec une précision significative (Oztekin, et al., 2016).

La force des réseaux réside dans leur capacité à apprendre et à discerner des schémas subtils et les liens linéaires et non linéaires dans un grand nombre de variables à la fois, sans être étouffés par les détails, et sans supposer aucune hypothèse sur les variables. Ils détectent les dépendances à partir des échantillons exemple. Ils peuvent également effectuer plusieurs opérations simultanément. Ils peuvent non seulement identifier les liens dans quelques variables, mais également détecter des relations dans des centaines de variables. Même lorsqu'un ensemble de données est bruyant ou comporte des entrées non pertinentes, les réseaux

peuvent apprendre des caractéristiques importantes des données. Les entrées qui peuvent sembler non pertinentes peuvent en fait contenir des informations utiles. Ils peuvent également s'adapter au changement de comportement du marché. Ce sont ces caractéristiques qui font particulièrement la force des réseaux de neurones artificiels.

Beaucoup de recherches ont montré que les prévisions basées sur les réseaux de neurones sont meilleures, car ils sont capables de capter les relations non-linéaires entre les variables et ne supposent aucune hypothèse sur les données. Bahrammirzaee's (2010) a détaillé dans sa recherche les différents articles qui traitent le pouvoir prédictif des séries temporelles par les réseaux de neurones. Il a trouvé, d'après ces articles, que la technique des réseaux de neurones surperforme toujours dans la prévision de rendement des actions (Bahrammirzaee, 2010).

Olson & Mossman (2003) ont utilisé les données de 18 ans (1976-1993) pour comparer la technique des réseaux de neurones par rapport à une régression linéaire et la régression logistique. Ils ont expliqué le rendement ajusté par 61 ratios comptables comme étant des variables explicatives (ratio courant, ratio dette sur actifs permanents, ratio rendement sur les actifs, etc.). Ils ont trouvé que le modèle basé sur les réseaux de neurones artificiels surperforme en prévision les autres modèles traditionnels (Olson & Mossman, 2003).

Kumar (2009) a comparé la performance de deux modèles pour prévoir le rendement de l'indice boursier S&P500 aux États-Unis et l'indice HIS en Hong Kong en utilisant les rendements historiques comme des variables indépendantes, et ce, sur une période allant de l'année 1928 jusqu'à l'année 2008. Les deux modèles utilisés sont : 1) les réseaux de neurones artificiels (ANN) configurés en MLP avec une seule couche cachée et une fonction d'activation « Sigmoidale »; et 2) le modèle ARIMA. Ce dernier a été identifié en ARIMA (2, 1, 2) et ARIMA (1, 1, 1) tel qu'il est suggéré par la minimisation des deux critères AIC et BIC. Ils ont trouvé que les ANN ont plus de pouvoir prédictif que le modèle ARIMA (Kumar, 2009).

Niaki et Hoseinzade (2013) ont testé l'apport des réseaux de neurones artificiels dans la prévision des séries temporelles. Ils ont simulé une stratégie d'investissement portant sur les données quotidiennes de l'indice boursier américain S&P500 entre l'année 1994 et l'année 2008. Cette stratégie consiste à acheter l'indice quand le modèle prévoit une tendance haussière de celui-ci, et le vendre quand le modèle prévoit une tendance baissière. Cette stratégie a surperformé le rendement de l'indice durant la période étudiée. Le modèle utilise les « feed-forward » ANN pour expliquer la variable dépendante qui est le rendement de l'indice S&P500

par des vingtaines de variables indépendantes. Ces derniers incluent des indicateurs techniques, fondamentaux, financiers et économiques. Parallèlement, les auteurs comparent la prévision basée sur les ANN par rapport à celle basée sur la régression logistique, et ils ont conclu que la première prime sur la deuxième du fait que la relation entre la variable à expliquer et les variables explicatives est non-linéaire (Niaki & Hoseinzade, 2013).

6.3. Régression logistique (RL)

La régression logistique est utilisée tant pour la prédiction que pour la détermination de l'intensité de la relation entre les variables. Elle est utilisée quand la relation entre les variables est linéaire, elle est très bénéfique pour les modèles qui expliquent une variable dépendante dichotomique par des variables explicatives (prédicteurs) continues ou catégoriques. La régression logistique prédit la vraisemblance du résultat, sous forme d'une probabilité de survenance de l'évènement étudié, en se basant sur des variables indépendantes qui peuvent être de nature différente (Dutta, et al., 2012).

Les coefficients dans la régression logistique sont estimés par le maximum de vraisemblance. Ils peuvent être interprétés comme des odd-ratios des variables indépendantes. La RL modélise la probabilité d'occurrence d'un « succès » parmi les deux classes dichotomiques. La combinaison linéaire des prédicteurs est utilisée pour s'ajuster à la transformation « logit » des probabilités de « succès ». Le classement est assigné à la classe « succès » si la probabilité estimée est supérieure à un seuil préfixé (généralement 0.5). L'avantage de la RL est qu'elle applique une fonction de lien au modèle de régression linéaire, et que les variables indépendantes peuvent être dichotomiques, continues ou discrètes ou mélange des trois. Contrairement à la régression linéaire, l'hypothèse de la normalité des variables dans le modèle de la RL n'est pas nécessaire (Lee, 2004).

Altman (1968) et Ohlson (1980) étaient les pionniers à utiliser la régression logistique dans un modèle de prédiction. Ohlson (1980) se concentre sur la capacité du modèle à prédire le défaut des compagnies en fonction de leur probabilité de défaut. Après lui, plusieurs auteurs, comme Zavgren (1985) et Zmijewski (1984), le suivent dans l'utilisation de la RL pour prédire la faillite et le stress financiers des compagnies (Ohlson, 1980; Zavgren, 1985; Zmijewski, 1984).

Dutta et al. (2012) ont développé un modèle de régression logistique qui prédit la probabilité si l'action d'une compagnie est « good » ou « poor ». Ils ont utilisé dans leur modèle les

données des ratios financiers de 30 grandes firmes, dont les actions les plus activement échangées en bourse (Indian Stock Exchange) sur une période de quatre années (2005-2008). L'action est qualifiée « good » quand le rendement de l'action dépasse le rendement du marché (NIFTY Indian Index) et vice versa (Dutta, et al., 2012).

Carol Anne Hargreaves et al. (2013) ont comparé l'application de la régression logistique et les réseaux de neurones à perceptron multicouches (deux couches cachées) à retro-propagation pour construire un portefeuille d'actions et la mise en place d'une stratégie d'investissement à court terme (20 jours). Les deux techniques permettent la sélection de six titres qui ont la meilleure performance espérée parmi les 100-200 titres faisant partie aux deux secteurs, Santé et Finance, de l'indice boursier australien d'actions. Plusieurs indicateurs techniques et fondamentaux sont ordonnés en fonction de leurs importances selon l'algorithme d'importance des forêts aléatoires (Random Forest Importance Algorithm, Beirman 2001) qui affecte un score aux variables et les classe selon leur puissance prédictive de la variable dépendante. Dix variables ont été retenues après avoir testé plusieurs combinaisons de variables selon leurs degrés d'importance pour trouver le nombre optimal de variables offrant la meilleure performance selon le secteur. Le moment d'achat et de vente des titres (timing) est défini selon l'indice de force relative (RSI) et la performance du modèle est mesurée par le rendement relatif au rendement du benchmark (indice sectoriel et indice global). Les résultats ont montré que les deux techniques ont dégagé un rendement supérieur au benchmark (Hargreaves, 2013).

6.4. Les forêts aléatoires (RF)

Le RF est un algorithme très efficace et très populaire, il a été introduit par Leo Breiman pour pallier les problèmes de sur-apprentissage des arbres de décisions, il peut être appliqué tant aux problèmes de régression qu'aux problèmes de classification. De même, les variables prédictives peuvent être catégorielles ou continues. Le RF est un ensemble d'arbres de décision dont la prédiction est basée sur le vote de majorité en classification, et la moyenne dans le cas de la régression. En effet, le RF utilise les arbres de décision pour le paramétrage du modèle, et utilise l'approche ensembliste pour optimiser le modèle construit. Le RF utilise la technique de rééchantillonnage (Bootstrap) de telle sorte que chaque arbre de décision dans la forêt s'entraîne sur : 1) une partie de données échantillonnée aléatoirement avec remise selon la méthode « Bagging »; et 2) une partie de variables selon ce qu'on appelle « feature sampling ». Par cette technique, le RF empêche reproduire les mêmes arbres et permet aussi aux

observations mal représentées d'être plus visibles (Biernat & Lutz, 2016; Imandoust, & Bolandraftar, 2014).

Le RF fait partie des techniques ensemblistes, il est attrayant, car il possède les caractéristiques suivantes : 1) il gère à la fois la régression et la classification; 2) il est rapide tant en entraînement qu'en prédiction et ayant la possibilité de se mettre en œuvre en parallèle, car les arbres peuvent être construites indépendamment; 3) il a moins de paramètres de configuration; 4) il peut être utilisé directement pour des problèmes de grande dimension; 5) il mesure l'importance des variables; 6) il détecte les variables aberrantes, etc. (Zhang & Ma, 2012).

Li et Sun (2011) ont observé que les classificateurs multiples surperforment les classificateurs simples dans la précision de la prédiction et dans la rentabilité des investissements. Ils ont constaté qu'il n'y a pas de différence significative entre « vote par majorité » et « Bagging » dans la précision de la prédiction (Sun, Liao & Li 2013).

Booth A. et al. (2014) ont développé un système automatique pour prédire le rendement des actions durant les régularités saisonnières dans le prix des actions. Le système est composé d'un ensemble de RF (ensemble d'ensembles) pondéré en fonction de leur performance pour produire un signal d'achat ou de vente sous contrainte de risque de gestion (maximum drawdown). Quand un nouvel entraînement sur les données non encore vues aura lieu, il va s'ajouter à l'ensemble d'une manière en ligne pour permettre à l'algorithme de s'adapter aux changements de régime du marché. Ils ont utilisé les données de l'indice boursier allemand d'actions DAX entre l'année 2000 et l'année 2008 comme un échantillon d'entraînement, les données entre l'année 2008 et l'année 2010 comme étant le jeu de données de la validation croisée et finalement l'échantillon test pour les données entre l'année 2010 et l'année 2012. Les variables utilisées lors de cette étude sont des indicateurs techniques issus du prix d'ouverture, prix de clôture, prix haut et le prix bas. 23 prédicteurs ont été choisis parmi plusieurs selon leurs impacts sur la performance prédictive du modèle en adoptant la méthode de classement d'importance de composantes (feature importance ranking) proposée par Breiman (2001) (Booth, et al., 2014).

Khaidem et al. (2016) ont prédit la variation de prix des actions de quelques compagnies (Apple, Samsung, GE) sur les horizons d'un, deux et trois mois en appliquant l'algorithme ensembliste des forêts aléatoires. Les auteurs ont constaté que le problème de la prédiction des actions performe mieux quand il est traité comme étant un problème de classification (sens de

la variation des prix sur un horizon donné) au lieu d'un problème de régression, car la prédiction de la valeur exacte des prix d'actions est une tâche difficile à cause de leur nature chaotique et très volatile. Dans cette recherche, les auteurs ont calculé des indicateurs techniques (RSI, %K, Willaims %R, MACD, PROC, OBV) sur la série des prix exponentiellement lissée pour enlever le bruit des données historiques et de permettre au modèle d'identifier facilement la tendance des prix à long terme. Ils ont utilisé l'«Accuracy», la «Précision», le «Recall», la «Spécificité» et le ROC (Receiver Operationg Characterictic) pour mesurer la robustesse du modèle et ils ont abouti finalement que leur modèle est plus performant comparativement à d'autres modèles, comme SVM et les ANN, utilisés dans la littérature pour le même type de données (Khaidem, Saha & Dey, 2016).

6.5. Modèles hybrides

Selon Zhang (2003), les problèmes dans le monde réel ne sont pas strictement linéaires ou non-linéaires, pour capter la linéarité et la non-linéarité, il a comparé entre les trois modèles suivants : 1) le modèle linéaire ARIMA; 2) le modèle non-linéaire des réseaux de neurones artificiels (ANN); et 3) un modèle hybride entre ANN et ARIMA. Il a utilisé le cas simple du modèle ARIMA qu'est la marche aléatoire (Random Walk) pour expliquer la partie linéaire des données, et les ANN pour expliquer la partie résiduelle non linéaire du modèle. Après avoir entraîné son modèle sur les données de taux de change (Britick pound / US entre 1980 et 1993) et autres, il a obtenu que le modèle basé sur les ANN et le modèle hybride surperforme en précision le modèle linéaire basé sur la marche aléatoire, tandis que le modèle hybride surperforme les deux autres modèles séparément (Zhang, 2003).

En utilisant la même base de données de Zhang (2003), et les mêmes modèles (ANN avec configuration différente), Khashei et Bijari (2010) ont confirmé le résultat de Zhang (2003) que le modèle hybride en ANN et ARIMA offre plus de performance en prévision comparativement à ANN. Cette dernière performe également mieux qu'ARIMA (particulièrement la marche aléatoire) (Khashei & Bijari, 2010).

Maia et Carvalho (2011) ont combiné les ANN avec le lissage exponentiel de Holt's. Les ANN dans ce modèle sont configurées comme étant un perceptron multicouche à une couche cachée et une couche de sortie. Ils ont appliqué leur modèle sur le prix des actions de 15 firmes de différents secteurs d'activités, et ils ont trouvé que le modèle hybride est mieux qu'un modèle simple (Maia & de Carvalho, 2011).

Sharmila V.J. et Ramaswami M. (2017) ont combiné deux techniques d'apprentissage automatique, SVM et ANN, pour prédire la tendance des actions de six grandes compagnies de l'indice boursier NSE en Inde (National Stock Exchange of India) en utilisant les données quotidiennes de leurs prix d'actions entre janvier 2012 et décembre 2015. Les auteurs ont utilisé SVM pour la réduction et la sélection des variables pertinentes à l'étude qui expliquent le mieux la variable cible. Par cette technique, les variables qui introduisent beaucoup de bruit et de la redondance sont exclues. Ensuite, ils prévoient la tendance des prix d'action des compagnies choisies en appliquant aux variables réduites la technique ANN. Dans cette étude, 22 variables de l'analyse technique ont été utilisées comme des prédicteurs, à savoir : les moyennes mobiles simples, les moyennes mobiles exponentielles, les moyennes mobiles ajustées par le volume, Gamme de Pourcentage Williams (%R), l'indice des canaux de matières premières (Commodity Channel Index), les bandes de Bollinger, etc (Vaiz & Ramaswami, 2016).

Conclusion

Depuis l'apparition des marchés financiers structurés, la question sur l'efficacité des marchés et la prévisibilité du rendement d'actions étaient au centre de débat entre les différents courants de pensée. Une gestion active basée sur l'utilisation et la modélisation des données montre sa capacité à sélectionner des actions et à construire des portefeuilles plus profitables. Cette rentabilité excédentaire remet en cause la notion d'efficacité des marchés. D'ailleurs, c'est la raison pour laquelle, les agents financiers investissent en infrastructure logique et physique pour collecter les données.

Différents modèles ont été appliqués aux données collectées pour tenter d'expliquer et prédire le comportement des actions en vue de construire des portefeuilles surperformant le marché. Entre autres, les modèles classiques statistiques utilisent moins de données et assument des hypothèses statistiques sur la distribution des données. Toutefois, avec l'essor des nouvelles technologies et particulièrement les technologies de l'informatique, les données financières se sont produites et se sont accumulées à un rythme sans précédent. Face à la complexité de ces données, la quantité et la vitesse de leur production, les modèles classiques montrent leur incapacité à en tirer profit. D'où, la nécessité d'approches automatisées qui utilisent efficacement une quantité massive de données financières pour aider les investisseurs à la prise de décision d'investissement.

Il ressort de la revue de littérature ci-dessus que la plupart des recherches montrent que les modèles de type « data-mining » surperforment les modèles classiques de sélection de bons titres et de la prédiction de la meilleure performance des actions.

BIBLIOGRAPHIE

- Ankam, V. (2016). *Big data analytics with spark and hadoop*. Packt Publishing Limited.
- Bahrammirzaee, A. (2010). A Comparative Survey of Artificial Intelligence Applications in Finance : Artificial Neural Networks, Expert System and Hybrid Intelligent Systems. *International Journal of Neural Computing and Application*, Available online 20 June 2010.
- Biernat, É., & Lutz, M. (2016). *Data science fondamentaux et études de cas : Machine learning avec Python et R*. Eyrolles.
- Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, 81(3), 637-654.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307-327.
- Booth, A., Gerding, E., & McGroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, 41(8), 3651-3661.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting and control* Holden-Day (San Francisco).
- Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1), 57-82.
- Dutta, A., Bandopadhyay, G., & Sengupta, S. (2012). *Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression*. 7(1), 32.
- Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4), 987-1007. JSTOR.
- F., A., Elsir, S., & Faris, H. (2015). A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index. *International Journal of Advanced Research in Artificial Intelligence*, 4(7).
- Fama, E. F. (1970). Efficient Capital Markets : A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), 383.
- Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns On Stocks

And Bonds. *Journal of Financial Economics*, 33, 3–56.

- Fama, E. F., & French, K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, 51(1), 55-84.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.
- Genest, C., Ghouli, K., & Rémillard, B. (2007). Rank-Based Extensions of the Brock, Dechert, and Scheinkman Test. *Journal of the American Statistical Association*, 102(480), 1363-1376.
- Gómez-Ramos, E., & Venegas-Martínez, F. (2013). A Review of Artificial Neural Networks : How Well Do They Perform in Forecasting Time Series? *Analitika*, 6(2), 7-15.
- Grossman, S. (1976). On the Efficiency of Competitive Stock Markets Where Trades Have Diverse Information. *The Journal of Finance*, 31(2), 573-585. JSTOR.
- Grossman, S. J., & Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, 70(3), 393-408.
- Ehsan, H., Hamed, D. A., & Jamal, S. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, 2(7), 109-118.
- Hargreaves, C. A. (2013). Stock Portfolio Selection using Data Mining Approach. *IOSR Journal of Engineering*, 3(11), 42-48.
- Imandoust, S. B., & Bolandraftar, M. (2014). Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange. *International Journal of Engineering Research and Applications*, 4(6), 106-117.
- Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *The Journal of Finance*, 45(3), 881-898.
- Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1), 65-91.
- Kamley, S., Jaloree, S., & Thakur, R. S. (2016). Performance Forecasting of Share Market using Machine Learning Techniques : A Review. *International Journal of Electrical and Computer Engineering (IJECE)*, 6(6), 3196.
- Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market

prices using random forest. *ArXiv:1605.00003 [Cs]*.

- Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Syst. Appl.*, 37, 479-489.
- Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2), 307-319.
- Kumar, M. (2009). Nonlinear prediction of the Standard & Poor's 500 and the Hang Seng index under a dynamic increasing sample. *Asian Academy of Management Journal of Accounting & Finance*, 5(2).
- Lee, S. (2004). Application of Likelihood Ratio and Logistic Regression Models to Landslide Susceptibility Mapping Using GIS. *Environmental management*, 34, 223-232.
- Levy, R. A. (1967). Relative Strength as a Criterion for Investment Selection. *The Journal of Finance*, 22(4), 595-610. JSTOR.
- Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5), 15-29.
- Mahajan, K. S., & Kulkarni, R. V. (2013). A Review: Application of Datamining Tools For Stock Market. *International Journal of Computer Technology and Applications*, 4(1), 19.
- Maia, A. L. S., & de Carvalho, F. de A. T. (2011). Holt's exponential smoothing and neural network models for forecasting interval-valued time series. *International Journal of Forecasting*, 27(3), 740-759.
- Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1), 77-91.
- Merton, R. C. (1974). On the Pricing of Corporate Debt : The Risk Structure of Interest Rates*. *The Journal of Finance*, 29(2), 449-470.
- N., J., & Wold, H. (1939). A Study in Analysis of Stationary Time Series. *Journal of the Royal Statistical Society*, 102(2), 295.
- Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, 9(1), 1.
- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109.
- Olson, D., & Mossman, C. (2003). Neural network forecasts of Canadian stock returns

using accounting ratios. *International Journal of Forecasting*, 19(3), 453-465.

- Osborne, M. F. M. (1959). Brownian Motion in the Stock Market. *Operations Research*, 7(2), 145-173.
- O'Shaughnessy, J. P. (1997). *What works on Wall Street: A guide to the best-performing investment strategies of all time*. McGraw-Hill.
- Ouahilal, M., Mohajir, M. E., Chahhou, M., & Mohajir, B. E. E. (2017). A novel hybrid model based on Hodrick–Prescott filter and support vector regression algorithm for optimizing stock market price prediction. *Journal of Big Data*, 4(1), 31.
- Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research*, 253(3), 697-710.
- Patil, D. J. (2011). *Building data science teams: The skills, tools and perspectives behind great data science groups*. O'Reilly.
- Rather, A. M., Sastry, V. N., & Agarwal, A. (2017). Stock market prediction and Portfolio selection models: A survey. *OPSEARCH*, 54(3), 558-579.
- Roberts, H. (1967). "Statistical versus clinical prediction of the stock market", unpublished manuscript. Chicago, University of Chicago, Centre for Research on Security Prices.
- Roberts, H. V. (1959). Stock-Market "Patterns" and Financial Analysis: Methodological Suggestions. *The Journal of Finance*, 14(1), 1-10.
- Rosillo, R., Giner, J., & De la Fuente, D. (2014). Stock Market Simulation Using Support Vector Machines: Stock Market Simulation Using Support Vector Machines. *Journal of Forecasting*, 33(6), 488-500.
- Rouwenhorst, K. G. (1998). International Momentum Strategies. *The Journal of Finance*, 53(1), 267-284.
- Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 19(3), 425.
- Sun, J., Liao, B., & Li, H. (2013). AdaBoost and bagging ensemble approaches with neural network as base learner for financial distress prediction of Chinese construction and real estate companies. *Recent patents on computer science*, 6(1), 47-59.
- Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst*, 36, 1-12.

- Swamynathan, M. (2017). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Apress.
- Swinkels, L. (2004). Momentum investing : A survey. *Journal of Asset Management*, 5(2), 120-143.
- Taylor, H. M., & Karlin, S. (2014). *An Introduction to Stochastic Modeling*. Academic Press.
- Vaiz, J. S., & Ramaswami, M. (2016). A Hybrid Model to Forecast Stock Trend Using Support Vector Machine and Neural Networks. *International Journal of Engineering Research and Development (IJERD) Volume, 13*, 52-59.
- Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, 100, 55-63.
- Yule, G. U. (1926). Why do we Sometimes get Nonsense-Correlations between Time-Series? A Study in Sampling and the Nature of Time-Series. *Journal of the Royal Statistical Society*, 89(1), 1.
- Zavgren, C. V. (1985). Assessing the Vulnerability to Failure of American Industrial Firms : A Logistic Analysis. *Journal of Business Finance & Accounting*, 12(1), 19-45.
- Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media.
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.
- Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 22, 59-82. JSTOR.