Optimisation d'un logiciel de prédiction d'un indice boursier

Benjamin Jurczak
Telecom Nancy
Computer Science
Nancy, France
benjamin.jurczak@telecomnancy.eu

Thomas Balduz
Telecom Nancy
Economics, Computer Science
Nancy, France
thomas.balduz@telecomnancy.eu

Alexandre Dhenin
Telecom Nancy
Computer Science, Financial Market
Nancy, France
alexandre.dhenin@telecomnancy.eu

Mario Lezoche
MdC HDR
CRAN
Nancy, France
mario.lezoche@univ-lorraine.fr

***Abstract*—Cet article scientifique a pour vocation de faire comprendre au lecteur la démarche de prédiction du cours de la bourse, au travers notamment d'algorithmes d'intelligence artificielle (IA).**
Initialement, ce projet avait pour but l'optimisation d'un algorithme existant de prédiction de l'indice boursier allemand (DAX), élaboré par un précédent élève de deuxième année, Dumitru Bulgaru. Cependant, la documentation et les recherches menées nous ont montré que son (ses) algorithme(s) de prédiction étai(en)t déjà d'une très grande qualité et exploitable(s) en l'état, faisant de son travail notre point de départ.
Ce document présente les techniques mises en oeuvre pour tenter de créer des algorithmes encore meilleurs en terme de précision ainsi que les applications qui ont été trouvées à ces algorithmes de prédiction. Enfin, le but "ultime" était de pouvoir proposer une interface graphique permettant à l'utilisateur de réaliser son plan d'investissement pour gagner de l'argent grâce aux algorithmes sous-jacents.
Mots-clés –
- **DAX, bourse**
- **Prédiction, anticipation**
- **réseaux de neurones, random forests, régression linéaire**
- **séries temporelles, mouvements browniens**

I. Introduction

Prédire avec précision le cours de la bourse est un sujet qui fascine les boursicoteurs depuis toujours, sujet dont les enjeux ont été démultipliés avec l'arrivée de nouvelles pratiques d'investissements telles que le trading haute fréquence. Les récentes avancées des modèles de prédictions et notamment de ceux mettant en jeu l'intelligence artificielle ont permis des résultats inégalés jusqu'à présent. C'est dans ce contexte que le projet proposé prend du sens: chercher des moyens d'optimiser les algorithmes d'IA proposés. Bien qu'il existe des méthodes de prédiction, la tâche est rude étant donné la volatitilé des marchés financiers. Pour notre projet, nous nous sommes intéréssé au DAX car c'est un indice boursier possédant une faible volatilité.

Les modèles d'IA les plus courants sont en autres les réseaux de neurones artificiels, les k-voisins ou encore les random forest (forêts aléatoires en français) qui sont un peu plus élaborées. Toutes ces méthodes ont été utilisées dans ce projet pour réaliser des prédictions. Notre démarche a été dans un premier temps d'utiliser une combinaison linéaire des prédictions à notre disposition afin de maximiser le coefficient de détermination. Nous avons également regardé l'intérêt d'introduire le modèle de Gordon et Shapiro à cette pondération. Nous avons également étudié les séries temporelle adin d'obtenir un intervalle de confiance sur nos prédictions. Nous nous sommes ensuite intéressés aux mouvements browniens, l'hypothèse des marchés financiers efficients nous assure que les variations de prix sont essentiellement aléatoire et un processus stochastique continu pourrait alors permettre de modéliser ces variations.

II. État de l'art

Dans cette partie, nous donnerons au lecteur les connaissances nécessaires à la bonne compréhension du cadre projet et de son environnement (bourse, IA, ...), les domaines couverts étant complexes et nombreux.

A. Le marché des capitaux

Les marchés de capitaux représentent l'ensemble des marchés qui permettent d'échanger divers actifs financiers à long terme. Ils se subdivisent en trois parties : le marché financier, le marché obligataire et le marché monétaire. Les marchés de capitaux sont essentiels au fonctionnement économique d'un pays, en ce qu'ils remplissent deux fonctions primordiales : premièrement, les marchés de capitaux permettent la rencontre entre les agents économiques ayant un excédent de capitaux, et ceux ayant besoin de financement ; deuxièmement, les marchés de capitaux permettent aux sociétés d'investir dans des instruments financiers, dont les actions. Ainsi, les marchés de capitaux servent de variables de décision pour les investisseurs, plus précisément le marché financier.

B. Le marché financier

Le marché financier est le lieu économique où les personnes physiques, les entreprises privées et les institutions publiques peuvent s'échanger des capitaux au sens large. Il peut s'agir d'actifs financiers, de matières premières ou d'autres produits. On y retrouve le marché des actions, des obligations, des devises et de certaines marchandises. La bourse est l'un de ces marchés.

La bourse est une entité économique dans laquelle peuvent s'échanger divers actifs de façon standardisée (nombreuses règles et codes) à un prix donné. La bourse est donc un marché réglementé qui permet de faciliter les échanges. Elle offre une garantie aux acheteurs et aux vendeurs en cas de défaut de l'une des parties. Le prix des actifs est déterminé en fonction de l'offre et de la demande. La bourse est généralement dotée d'un régulateur qui s'assure qu'il n'y ait pas de fraude et une libre circulation des informations concernant les actifs (afin de ne pas biaiser le marché).

Un indice boursier, ou indice de prix des actions, est un indicateur de l'évolution des prix des actions. Il est constitué d'un ensemble d'actions relatives à un secteur économique ou une zone géographique comme un pays. Il s'agit d'un outil fondamental pour l'investisseur qui souhaite investir sur les marchés des capitaux, en particulier la bourse. Les indices boursiers les plus connues sont le CAC40, le Nasdaq ou encore le Dow Jones.

Notre présent projet porte sur le Deutscher AktienindeX (DAX) qui est le principal indice boursier allemand. Son portefeuille d'actions est composé des 40 plus importantes entreprises cotées à la Bourse de Francfort. Historiquement le DAX présente la volatilité la plus faible et est donc un indice idéal pour l'application d'algorithmes de prédiction.

Fig. 1. Salle de marché à la bourse de Francfort

Les actions sont l'un des instruments les plus populaires sur le marché financier. Cela s'explique par le fait que les actions peuvent offrir un taux de rendement élevé assez rapidement. Cependant, la conséquence de ce taux élevé est un taux de risque important. La fluctuation du prix des actions affecte la décision de l'investisseur d'investir plus ou moins son capital. Une erreur lors de la prise de décision entraînera des pertes sur le long terme pour l'investisseur. Ainsi, pour minimiser le risque élevé du marché des actions, l'investisseur a besoin d'informations fiables pour prendre des décisions sur les actions qu'il doit acheter, conserver, ou vendre.

Deux facteurs ont une influence significative sur la modélisation du prix des actions, à savoir l'état antérieur de l'action qui va influencer le prix actuel et la réponse de l'action à l'actualité économique. On peut rajouter à ces facteurs l'hypothèse de l'efficience des marchés. L'hypothèse de l'efficience des marchés financiers veut que le prix d'un actif financier soit égal à sa valeur intrinsèque (fondamentale). C'est à dire qu'à un instant donné le prix de l'actif est tel qu'il prend en considération toutes les informations disponible le concernant. La conséquence de cela est qu'il n'y a pas d'arbitrage possible, le prix est fondamentalement correct et on ne peut ainsi "battre la bourse". Les variations futures du prix de l'actif sont le fait d'évènements qui ne se sont pas encore produits et sont donc aléatoires. On dit que ces variations suivent un processus de Markov.

C. Le modèle de Gordon et Shapiro

Le modèle de Gordon et Shapiro est un modèle d'actualisation des flux futurs [4]. Considérons une action, soit D_i l'anticipation à l'année 0 du dividende dans i années, P_i l'anticipation à l'année 0 du prix de l'action dans i années.

La rentabilité exigée de l'action est égale au rendement espéré + taux de plus-value espérée.
Soit $r_{cp} = \frac{D_1}{P_0} + \frac{P_1 - P_0}{P_0}$

On réécrit, $P_0 = \frac{D_1 + P_1}{1 + r_{cp}}$

On a de même, $r_{cp} = \frac{D_2}{P_1} + \frac{P_2 - P_1}{P_1}$

Puis, $P_1 = \frac{D_2 + P_2}{1 + r_{cp}}$

Ainsi, par récurrence immédiate, $P_0 = \sum_{i=1}^{\infty} \frac{D_i}{(1+r_{cp})^i}$

Avec l'hypothèse, $\lim_{i \to +\infty} \frac{P_i}{(1+r_{cp})^i} = 0$

On fait l'hypothèse que le taux de croissance du dividende est constant d'une année à l'autre, on a $D_i = D_0 * (1+g)^i$

Ainsi, $P_0 = \sum_{i=1}^{\infty} \frac{D_0 * (1+g)^i}{(1+r_{cp})^i} = \frac{D_0 * (1+g)}{r_{cp} - g} = \frac{D_1}{r_{cp} - g}$

On obtient alors le prix de l'actif financier en fonction des projections de dividende sur l'année suivante en supposant que le taux de croissance du dividende est constant à travers le temps.

D. Les séries temporelles

On appelle série temporelle [5], une séquence $\{Y_t\}_{-\infty}^{+\infty}$ de variables aléatoires indicées par le temps t

La série temporelle $\{Y_t\}$ est stationnaire en covariance si elle possède les propriétés suivantes :

$$\begin{cases} \forall t, E(Y_t^2) < +\infty \\ \forall t, E(Y_t) = \mu \\ \forall t, h, Cov(Y_t, Y_{t-h}) = \gamma_h \end{cases}$$

γ_h est appelée fonction d'autocovariance

Un bruit blanc gaussien est donné par
$Y_t = \epsilon_t \sim i.i.d.\mathcal{N}(0, \sigma_\epsilon^2)$

On définit l'opérateur de retard L tel que $LY_t = Y_{t-1}$

On définit les processus $ARMA(p, q)$ (modèle autorégressif et moyenne-mobile d'ordres (p,q)) ci dessous [5]:
$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + ... + \Phi_p Y_{t-p}$
$+ \epsilon_t$
$+ \Theta_1 \epsilon_{t-1} + \Theta_2 \epsilon_{t-2} + ... + \Theta_p \epsilon_{t-q}$
Soit, $Y_t(1 - \sum_{i=1}^{p} \Phi_p L^p) - \epsilon_t(\sum_{i=1}^{q} \Theta_q L^q) = \Phi_0 + \epsilon_t$

Ces processus représentent les séries temporelles comme des fonctions de leurs valeurs passées ainsi que les valeurs actuelles et passées de leur terme d'erreur. A noter que les processus ARMA sont utilisés pour modéliser les séries temporelles stationnaires. Le but est ainsi de déterminer quel processus $ARMA(p, q)$ modélise le mieux la série temporelle d'intérêt (le DAX dans notre cas). En se basant sur la méthode de Box et Jenkins [8], 4 étapes sont nécessaires: l'identification du modèle, l'estimation des paramètres du modèle, la validation de la procédure au moyen de test de diagnostic, la prévision de la série à l'aide du modèle sélectionné.

E. Outils statistiques

Une méthode doit pouvoir être "notée" afin de quantifier la qualité de sa prédiction. Pour cela on s'intéresse aux écarts entre les résultats que l'on obtient et la réalité (les résidus).

Le coefficient de détermination linéaire de Pearson, noté R^2, est une mesure de la qualité de la prédiction d'une régression linéaire. $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Lorsque la prédiction est bonne, ce coefficient tend vers 1.

Le MAE pour Mean Absolute Error est un coefficient qui est égal à la moyenne des résidus i.e. des erreurs de prédiction.
$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$

Ce coefficient tend vers 0 lorsque la prédiction tend vers la réalité.

Le MSE pour Mean Square Error est un coefficient analogue possédant les mêmes propriétés.
$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

F. Modèles d'IA principaux

1) Forêts aléatoires: Les forêts aléatoires [6] constituent un outil puissant d'apprentissage automatique reposant sur l'apprentissage par arbre de décision: une forêt aléatoire est en fait un ensemble d'arbres de décision. Le principe de fonctionnement est présenté sur la figure 2:

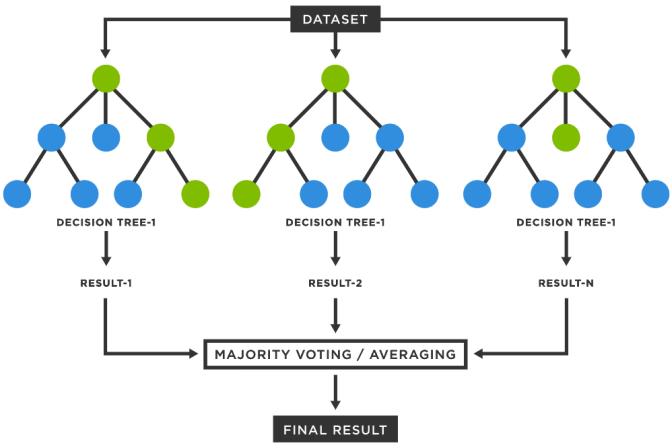

Fig. 2. Principe de fonctionnement des forêts aléatoires

A partir d'un jeu de données composées d'attributs, aussi appelés features, un certain nombre d'arbres de décisions sont créés. Les features sont des critères discriminants permettant de séparer le noeud initial (toutes les données) en plusieurs branches (clusters) pour grouper les données proches et les séparer des autres. Un arbre décisionnel cherche à répondre au problème suivant: classer une nouvelle donnée dans le bon cluster grâce à la séparation élaborée avec l'ensemble d'entraînement.
Le principe des randoms forests est d'avoir beaucoup d'arbres pour affiner la décision finale, d'où cette idée de forêt. Chacun de ces arbres est construit en utilisant des features aléatoires. Les randoms forests exploitent le principe d'intelligence collective: la meilleure décision à prendre est généralement celle qui est choisie par le plus grand nombre, ici par le plus d'arbres.

2) Réseaux de neurones artificiels: Un réseau de neurones [7] est un ensemble de couches de neurones mises en cascade, l'entrée d'un neurone étant la sortie des neurones de la couche précédente (exemple figure 3).

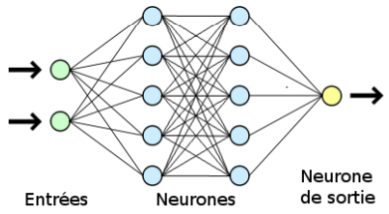

Fig. 3. Structure d'un réseau de neurone

Les informations échangées entre les neurones du réseau ne sont autres que des intensités: analogie avec les véritables neurones humains et de l'intensité électrique qui y circule. En effet, un neurone émet (et reçoit) une certaine intensité: souvent 0 ou 1 pour simplifier, le neurone est donc soit "actif" (1) soit "dormant" (0). Le fonctionnement d'un neurone du réseau est schématisé sur la figure 4 et est le suivant:

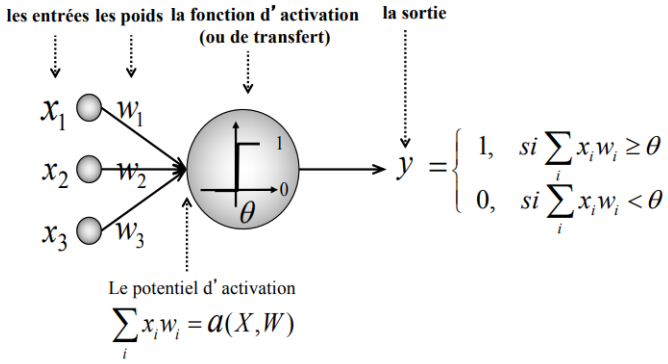

Fig. 4. Structure d'un neurone

Le neurone reçoit en entrée la somme des intensité des neurones de la couche précédente, chacune multipliée par le poids de l'arête associée. Cette entrée est envoyée dans la **fonction d'activation** du neurone. Soit l'entrée est suffisante pour activer la fonction (le neurone enverra 1), soit elle l'active pas (le neurone enverra 0). Le processus se répète alors avec les couches suivantes.
Note: La couche initiale reçoit quant à elle les données "brutes", mais le principe d'activation est le même

3) Régression linéaire: Principal outil de l'économétrie, la régression linéaire consiste à supposer l'existence d'une relation linéaire entre la variable à expliquer y (ici le cours de la bourse) et les autres paramètres (features vus plus haut). Ces paramètres notés $(x_1, x_2, ..., x_n)$ dans le modèle peuvent être nombreux et leur choix dépend de la situation. Le modèle de régression linéaire consiste à supposer [1] qu'il existe des coefficients $(\beta_0, \beta_1, ..., \beta_k)$ et une erreur ϵ tels que:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_n + \epsilon$$

Avec n observations de la forme précédente, on aurait donc:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_1, 1 + ... + \beta_k x_1, k + \epsilon_1 \\ ... \\ y_n = \beta_0 + \beta_1 x_n, 1 + ... + \beta_k x_n, k + \epsilon_n \end{cases}$$

Cela donne sous forme matricielle:

$$Y = X\beta + \epsilon \text{ où } Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ ... \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ ... \\ \epsilon_n \end{pmatrix}$$

Rappelons que l'on cherche à réaliser la meilleure prédiction possible. Cela revient à chercher les coefficients de la matrice β qui rendent les erreurs ϵ_i les plus petites possibles. La méthode d'estimation des *moindres carrés* montre que c'est le cas lorsque:

$$\beta = (X^T X)^{-1} X^T Y$$

La simplicité de la formule fait de la méthode de régression linéaire un très bon modèle de prédiction en terme de temps de calcul. Cependant, il faut choisir les paramètres qui vont rendre l'estimation la plus précise possible, paramètres présents par centaines voire milliers en économie.

G. Le mouvement brownien géométrique

D'après les hypothèses énoncées sur l'évolution du prix des actions, la prévision est la meilleure méthode pour prédire le prix futur de l'action. Cependant, ces prévisions présentent un taux de risque relativement élevé. De nombreuses recherches ont été menées en mathématiques financières afin de baisser ce taux et le mouvement brownien géométrique est aujourd'hui largement utilisé dans la modélisation du cours des actions. Il s'agit d'un processus stochastique standard (appelé aussi processus de Wiener) qui suppose que les profits ou pertes sur les actions sont indépendants entre eux et normalement distribués.

1) Construction de l'estimateur du prix de l'action:

En se basant sur les recherches et résultats de D. Bulgaru, nous avons l'équation différentielle stochastique suivante:

$$dS_t = \mu S t dt + \sigma S t dB_t$$

avec :

- St le prix l'action au temps t ;
- μ la moyenne de rendement ;
- σ la volatilité ;
- B_t le mouvement brownien avec drift.

On rappelle que $B_t = \mu t + \sigma W_t$, avec une espérance égale à $E(B_t) = \mu t$ et une variance égale à $Var(B_t) = \sigma^2 t$.

On en déduit, en utilisant le lemme d'Itô [3], la solution de l'équation différentielle stochastique :

$lnS_t = lnS_{t-1} + (\mu - \frac{\mu}{2})dt + \sigma\epsilon\sqrt{dt}$
D'où $S_t = S_{t-1}e^{(\mu - \frac{\sigma^2}{2})dt + \sigma\epsilon\sqrt{dt}}$

La solution ci-dessus est appelée modèle de mouvement brownien géométrique avec drift du futur prix de l'action S_t avec la moyenne égale à $lnS_{t-1} + (\mu - \frac{\sigma^2}{2})t$ et la variance égale à $\sigma^2 t$. Cette équation peut être dérivée à partir de la valeur initiale S_0 en appliquant la formule sur la période de temps t, ce qui donne :

$S_t = S_0 e^{(\mu - \frac{\sigma}{2})t + \sigma\epsilon\sqrt{t}}$

Il s'agit d'une variable aléatoire à distribution log-normale. La moyenne est égale à :

$E(S_t) = e^{lnS_0 + (\mu - \frac{\sigma^2}{2})t + \sigma^2 t}$

$= e^{lnS_0} e^{(\mu - \frac{\sigma^2}{2})t + \sigma^2 t}$

$= S_0 e^{(\mu + \frac{\sigma^2}{2})t})$

2) Calcul du taux de rendement:

Ici est présenté la version classique du taux de rendement entre l'instant t et t-1.

$R_t = \frac{S_t - S_{t-1}}{S_{t-1}}$
Avec :

- R_t : le taux de rendement à l'instant t ;
- St le prix de l'action à l'instant t ;
- $St - 1$ le prix de l'action à l'instant t-1.

3) Estimation de la valeur du drift et de la volatilité:

Le drift et la volatilité sont les deux paramètres à estimer. La méthode d'estimation est déterminante pour la qualité des prévisions. Nous commençons par étudier une estimation du drift et de la volatilité constante. Le drift est défini de la façon suivante :
$\mu = \frac{1}{M\delta t} \sum_{t=1}^{M} R_t$
Avec :

- R_t : le taux de rendement à l'instant t ;
- μ le drift;
- M le nombre de taux total.

Après avoir calculé le drift moyen, nous pouvons obtenir la valeur de la volatilité de deux façons : la volatilité moyenne classique et la volatilité logarithmique. Voici les équations correspondantes :

$\sigma_1 = \sqrt{\frac{1}{(M-1)\delta t} \sum_{t=1}^{M} (R_t - \bar{R})^2}$

$\sigma_2 = \sqrt{\frac{1}{(N-1)\delta t} \sum_{t=2}^{M} (logS_t - logS_{t-1})^2}$

Avec :

- σ_1 la volatilité ;
- σ_2 la volatilité logarithmique ;
- M le nombre de taux total ;
- N le nombre d'actions.

4) Intervalle de confiance:

Afin de tester la précision des prévisions du mouvement Brownien géométrique, nous allons implémenter un niveau de confiance à 95%. Nous partons de l'équation suivante :

$lnS_t = lnS_0 + (\sigma - 1/2\sigma^2)t + \sigma B_t$

Comme il a été vu précédemment, la moyenne du prix de l'action S_t est égale à $lnS_0 + (\mu - \frac{\sigma^2}{2})t$ et la variance à $\sigma^2 t$. Pour obtenir un prix de l'action avec une confiance à 95%, nous avons :

$lnS_0 + (\sigma - 1/2\sigma^2)t - 1,96\sigma\sqrt{t} \leq$ ln $S_t \leq$ ln $S_0 + (\sigma - 1/2\sigma^2)t + 1,96\sigma\sqrt{t}$

Soit :
$e^{lnS_0 + (\sigma - 1/2\sigma^2)t - 1,96\sigma\sqrt{t}} \leq S_t \leq e^{lnS_0 + (\sigma - 1/2\sigma^2)t + 1,96\sigma\sqrt{t}}$

Avec :

- S_0 le prix l'action de départ de l'action ;
- S_t le prix l'action au moment t ;
- μ le drift ;
- σ la volatilité ;

5) Prévision du prix de l'action:

A cette étape, le prix de l'action est prédit selon les estimateurs présentés plus haut. Nous choisissons de présenter deux prédicteurs utilisant le mouvement brownien géométrique : l'un avec la volatilité normale et l'autre avec la volatilité logartihmique. Les équations pour les prédicteurs 1 et 2 sont les suivantes : $F_t = S_{t-1} + e^{(\mu - \frac{\sigma^2}{2})t + \sigma B_t}$
Avec :

- F_t le prix l'action prédit à l'instant t ;
- S_{t-1} le vrai prix l'action au moment t-1 ;
- μ le drift ;
- σ la volatilité, avec $\sigma = \sigma_1$ ou $\sigma = \sigma_2$;
- $B_t = \epsilon\sqrt{t}$.

6) Précision des prédictions:

Afin de tester la prévision de nos prédictions, nous comparons la valeur prédite avec la valeur réelle en utilisant le coefficient MAE présenté plus haut. Dans ce cas-ci,

l'équation du MAE est la suivante :

$$MAE = \frac{\sum_{t=1}^{N} |S_t - F(S_t)|}{N}$$

Avec :

- S_t le prix réel de l'action à l'instant t ;
- $F(S_t)$ le prix prédit de l'action à l'instant t ;
- N le nombre de données.

III. Méthodologie et résultats

A. Principe de fonctionnement de la prédiction

Le principe de prédiction consiste à utiliser les informations passées pour en déduire ce qui a le plus de chance de se produire dans le futur. En effet, il est couramment admis que le prix d'une action à un instant donné dépend de ses prix passés. Ainsi, le fonctionnement est le suivant:
Pour un jour donné, on utilise les n derniers jours (en fait les données de volume, prix, augmentation, baisse ... de ces n derniers jours) pour prédire le prix dans k jours. $k = 1$ correspondant à demain, $k = 2$ pour le surlendemain etc...
On choisira le plus souvent (n,k) = (1,1) c'est-à-dire, utiliser les informations de la veille pour prédire le lendemain. Il faut répéter cette opération répéter pour obtenir les prédictions sur les jours suivants.
Suivant ce principe, on observe sur la figure 5 le résultat obtenu (par M. Bulgaru) avec un modèle de régression linaire. La courbe bleue (cours réel) et la courbe rouge (cours prédit) se superposent à première vue de manière quasiment parfaite (légers écart en réalité, la courbe rouge a une léger retard).

Fig. 5. Prédiction avec une régression linéaire

Le coefficient de détermination R2 est égal à 0.991335 en cela la prédiction est excellente. Notre travail consiste donc à améliorer cette prédiction.

B. Optimisation

Une première approche a été de maximiser le coefficient de détermination R2. Nous avons préalablement analysé les différentes approches proposées par Monsieur Bulgaru. Nous avons remarqué que seules les prédictions par régression linéaire, forêts aléatoires et réseaux de neurones artificiels apportaient des résultats probants. Nous nous sommes alors demandé si une combinaison linéaire des ces approches apporterait des améliorations substantielles. Pour cela nous avons procédé par force brute. Le but étant de tester toutes les combinaisons possibles et de regarder laquelle maximisait le coefficient de détermination. Avec cette méthode, nous avons obtenu un R2 de 0.991959 soit une amélioration de 0.06% avec 3.35 % de ANN et 96.65 % pour la régression linéaire. Cependant, ce résultat n'est pas convenable pour plusieurs raisons. La complexité de cet algorithme est en $O(n^4)$ (Toutes les combinaisons possible puis l'appel à la fonction de calcul du R2) donc le temps de calcul est relativement long. Les améliorations ne sont pas non plus probantes le MAE n'est amélioré que de 50 centimes soit une diminution de 0.64 %. En continuant cette approche de combinaison linéaire pour minimiser le MAE ou bien maximiser le pourcentage de bonne variation, nous n'avons pas obtenu de résultat satisfaisant.

C. Nouvelles approches

Afin d'améliorer la prédiction nous nous sommes orientés vers de nouvelles méthodes.

1) Introduction du modèle de Gordon et Shapiro:

Afin de réduire le MAE - révélateur de l'erreur moyenne du prix prédit, il vaut 106€ pour M. Bulgaru -, nous avons eu l'idée de lisser nos résultats. Le but étant de trouver un moyen de quantifier au mieux la valeur du DAX sur une période donnée avec des projections et des hypothèses assez larges afin que le coefficient de lissage soit pertinent sur l'ensemble de la période étudiée.

La décision a été d'appliquer le modèle de Gordon et Shapiro. Pour utiliser ce modèle, il a fallu effectuer des approximations. A partir de l'historique des dividendes des sociétés composant le DAX nous avons déterminé que sur les dernières années le taux moyen de dividende par action est de 0.6391. La croissance du taux de dividende d'une année à l'autre est d'en moyenne 2.32%. De plus, sur les dix dernières années, la rentabilité moyenne annuelle de l'indice s'établit à 8,2%.
Ce sont donc les données que nous utiliserons pour déterminer un prix de l'actif permettant de lisser les résultats d'une année à l'autre. Pour résumer, la rentabilité exigée est de 8.2%, la croissance des dividendes de 2.32% et le dividende versé à l'année 0 de 0.6391.
On détermine alors $P_0 = \frac{0.6391*(1+0.0232)}{0.082-0.0232} = 11.1212$. Ainsi, avec ces hypothèses, une action d'un cours composant le DAX a en moyenne une valeur de 11.1212€. Afin de calculer l'équivalent en terme de points (l'unité utilisé pour les indices boursiers), multiplions cette valeur par la pondération associée à chaque société du DAX. Le score obtenu est alors de 11043.3516.
Nous avons alors essayé d'intégrer ce coefficient à notre modèle d'optimisation décrit ci dessus. Notre modèle ignore la

méthode de Gordon et Shapiro (il lui applique une pondération nulle). L'utilisation d'un modèle d'actualisation des flux futur n'améliore donc pas notre prédiction.

2) Utilisation des processus ARMA(p,q):

Le but est donc de modéliser le DAX (qui est une série temporelle) à l'aide d'un processus $ARMA(p,q)$. A priori, si on se réfère à la figure 5, la série temporelle semble avoir une tendance à la hausse (ce qui donne un indice de non stationnarité). Pour s'en assurer on applique un test de Dicky-Fuller. Le test est égal à -1.96 donc supérieur à -2.9 ce qui implique que la série n'est pas stationnaire. Pour rendre cette série stationnaire nous allons faire sa différence i.e. $\forall t, Y_t = Y_t - Y_{t-1}$. Puis, appliquons de nouveau un test de Dicky-Fuller, la série est cette fois bien stationnaire.

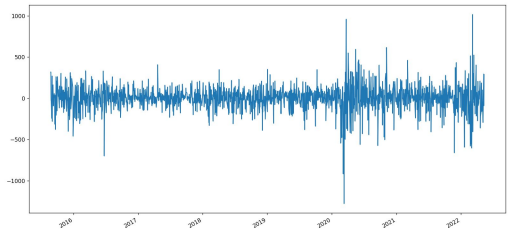

Fig. 6. DAX rendu stationnaire

La première étape est d'identifier notre modèle. Il faut alors chercher à minimiser le critère d'information d'Akaike (mesure de la qualité d'un modèle statistique). Pour cela, le parcours avec une boucle des combinaisons (p,q) permet de déterminer à quel moment le critère est minimal pour le modèle $ARMA(p,q)$ en question.
Pour le DAX, le critère est minimal pour $p = 3$ et $q = 2$.

$$Y_t(1 - \sum_{i=1}^{3} \Phi_p L^p) - \epsilon_t(\sum_{i=1}^{2} \Theta_q L^q) = \Phi_0 + \epsilon_t$$

Il faut désormais estimer les paramètres du modèle. Pour cela, nous utilisons la fonction SARIMAX de la librarie *statsmodel* et obtenons le modèle suivant:
$Y_t = 1.3778 + 1.8679Y_{t-1} + 0.9858Y_{t-2} + 0.0009Y_{t-3} + \epsilon_t + 1.8633\epsilon_{t-1} + 0.9741\epsilon_{t-2}$
Il faut ensuite valider la procédure via des tests de diagnostic.

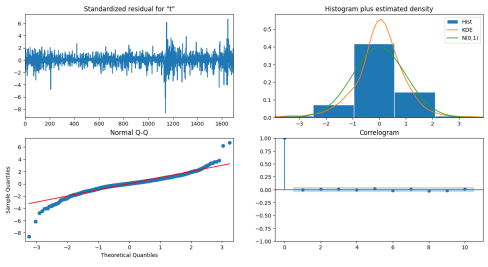

Fig. 7. Tests de diagnostic du modèle ARMA(3,2)

Les tests présents sur la figure 7 laissent raisonnablement penser que les résidus suivent bien une loi normale et donc que le modèle est cohérent.
Grâce au modèle, nous avons la possibilité de réaliser une prédiction d'un jour à l'autre en nous donnant un intervalle de confiance.

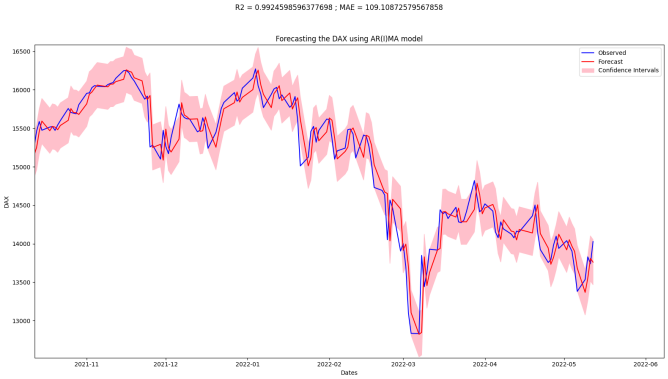

Fig. 8. Prédiction du DAX d'un jour à l'autre ; zoom sur le période Novembre 2021 - Mai 2022

Le modèle a donc un coefficient de détermination R2 environ égal à 0.9925 et un MAE de 109.109 (figure 8). Ce sont des résultats similaire à ceux obtenu par Monsieur Bulgaru (bien que sensiblement plus performant). L'intérêt de cette méthode réside essentiellement dans l'intervalle de confiance. En effet avoir une mesure de la fiabilité du système est nécessaire pour une application concrète de cet "outil" à des fins d'investissement.

Il est également possible de réaliser une prédiction dynamique c'est-à-dire sur une période futur donnée à partir des données existantes (figure 9). Une telle utilisation n'est pas judicieuse car le modèle reste centré autour de la moyenne et l'intervalle de confiance explose nécessairement.

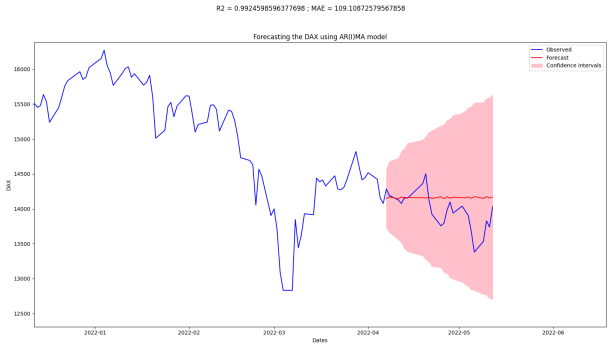

Fig. 9. Prédiction du DAX dynamiquement ; Prédiction sur une période de 25 jours

Les processus $ARMA(p,q)$ sont donc des outils puissants pour modéliser les séries temporelles stationnaires ayant une variance régulière. Ils sont efficace pour prédire l'avenir à

court terme mais à plus long terme, les erreurs ce cumulant, le modèle perd en fiabilité. Les actifs financiers sont souvent soumis à des chocs temporaires qui font que la volatilité est importante sur une courte période. Ce n'est pas le cas pour le DAX car il présente une faible volatilité. Cependant dans le cas où l'on voudrait modéliser un produit financier plus volatile, il faudrait se tourner vers un autre modèle. Les modèles $ARCH$ (AutoRegressive Conditional Heteroskedasticity) [5] semblent tout indiqués car pour déterminer l'erreur (qui était un bruit blanc Gaussien pour le modèle $ARMA$) ils prennent en compte l'erreur des termes précédent.

3) Mouvements Browniens:

A l'aide des équations établies plus haut, nous pouvons à présent établir des prévisions sur le prix futur de l'action. Tout d'abord, les données choisies pour établir nos prévisions à l'aide du mouvement brownien géométrique sont les prix de l'année 2020.
La grande volatilité en début d'année suivie d'une phase de croissance nous permet de confronter nos prédictions à des mouvements brusques du marché financier. Nous commençons par établir cinq prédictions du cours de l'action sur l'année 2020. Pour cela, nous commençons par calculer le drift et la volatilité du prix de l'action sur l'entièreté de l'année. Nous produisons ensuite cinq prévisions dont les tracés sont affichés figure 10.

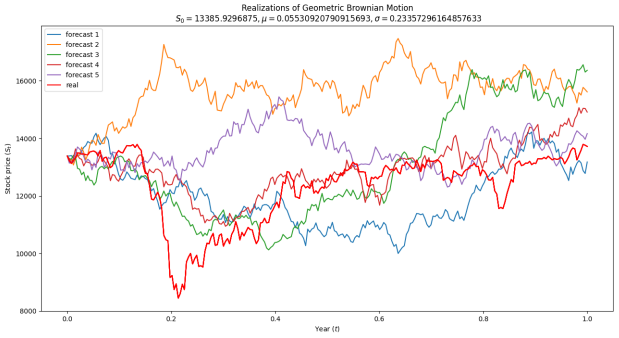

Fig. 10. 5 estimations browniennes sur l'année 2020

Avec le tracé de la courbe réelle en rouge, nous observons qu'aucune des estimations ne semble très pertinente. Afin de s'assurer que cette estimateur est capable d'estimer chacune des valeurs réelles du prix de l'action, nous procédons à mille estimations browniennes. Le résultat est le suivant :

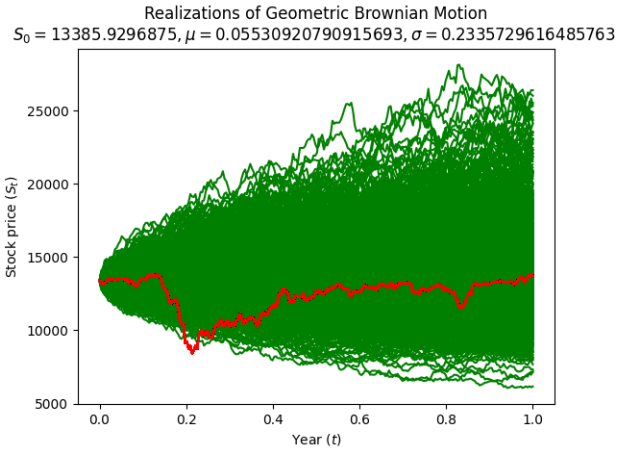

Fig. 11. 1000 estimations browniennes sur l'année 2020

Nous pouvons voir sur ce graphique que la courbe rouge est dans la zone verte, ce qui veut dire qu'elle est estimable par le mouvement brownien. Une légère nuance doit être apportée sur le fait qu'aucune des mille estimations n'a pu prédire le point le plus bas présent en x = 0.2, et qui correspond à l'arrivée du covid-19 en Europe. Afin d'améliorer notre modèle, nous décidons d'utiliser un drift flottant et une volatilité flottante. Ces deux paramètres, restés constants jusqu'alors, s'actualisent maintenant selon une fenêtre d'un certain nombre d'unités temporelles. L'idée serait donc de prédire le prix de l'action à un horizon d'une unité temporelle. Nous présentons ici le résultat de cette prédiction avec une fenêtre temporelle réglée sur un mois. Voici le résultat :

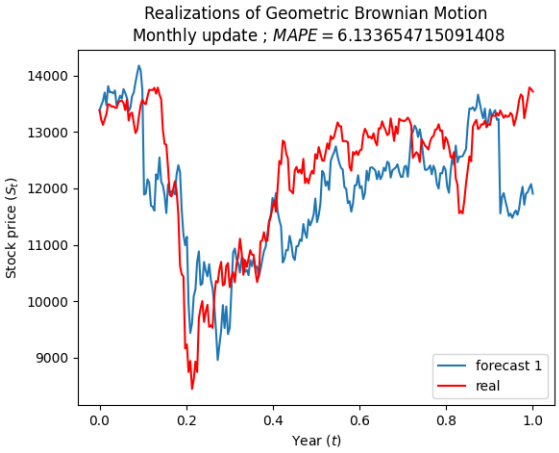

Fig. 12. Estimation brownienne avec une fenêtre temporelle fixée à 1 mois

Ici, le résultat se rapproche de la courbe réelle du prix de l'action. L'indicateur MAPE est égal à 6.1%, ce qui est acceptable mais pas encore suffisant. Après de multiples tentatives d'améliorations, nous avons choisi de changer de modèle de prédiction. En effet, les limites du mouvement brownien géométrique sont les suivantes :

- La prédiction du prix futur est aléatoire. Ce postulat n'est pas vrai, car les évolutions du prix dépendent du contexte économique ainsi que d'évènements décisifs comme la crise des sub-primes.
- Les fluctuations passées n'ont pas d'influence sur les fluctuations futures. Encore une fois, un contexte de croissance économique et de confiance dans les marchés va générer des tendances continues dans le temps.
- Les paramètres restent constants. Nous avons tenté de pallier cette limite avec la mise en place d'une fenêtre temporelle. D'autres méthodes existent comme le modèle d'élasticité de volatilité constante ou logarithmique [2], mais nous ne les avons pas implémentées.

IV. Conclusion

Pour conclure, ce projet de prédiction du cours du DAX s'est passé en deux temps avec d'abord une partie importante sur l'état de l'art, incontournable dans un projet de recherche, puis l'élaboration de méthodes d'optimisation des algorithmes initiaux.

Pour commencer, plusieurs idées ont été imaginées afin de tenter d'améliorer la précision et la confiance des prédictions. Dans un premier temps, il a été question d'utiliser les modèles les plus fiables (réseaux de neurones et régression linéaire) afin de les associer par combinaison linéaire pour en tirer les avantages de chacun. Cette méthode n'a pas été retenue à cause de l'amélioration minime qu'elle apportait par rapport à l'importance de sa complexité.

Ensuite, le modèle de Gordon et Shapiro s'est avéré être inefficace du fait des trop nombreuses approximations faites. L'approche par le modèle $ARMA(p, q)$ appliqué au DAX (qui est une série temporelle) est très efficace pour des prédictions à court terme. Les séries temporelles constituent une approche prometteuse. Cependant le modèle $ARMA$ n'étant fiable qu'à court terme, il est nécessaire d'utiliser d'autres méthodes/modèles pour des projections à moyen/long terme, le modèle $ARCH$ constitue une piste d'amélioration.

Parallèlement, le mouvement brownien géométrique a aussi été exploré. Cependant, notre modèle assez simpliste fait des hypothèses trop grossières, et ne permet donc pas d'améliorer nos précédents résultats. Ces limites peuvent néanmoins être affaiblies, notamment en approfondissant la recherche mathématique des processus stochastiques. Le mouvement brownien reste donc un modèle intéressant qui peut servir de support à des modèles plus complexes.

Enfin, ce projet aura permis de déceler les pistes d'améliorations qui présentent le plus de potentiel. Ainsi, l'approche qui semble la plus sérieuse à approfondir est celle de l'utilisation des séries temporelles éventuellement combinées avec un mouvement brownien.

References

[1] C. Anghelache, G. Anghel, L. Prodan, C. Sacala, and M. Popovoci. *Multiple Linear Regression Model Used in Economic Analyses (p. 120-125)*. PhD thesis, Academy of Economic Studies, Bucharest, 2014.

[2] Axel A. Araneda. The fractional and mixed-fractional cev model. *Frankfurt Institute for Advanced Studies*, June 1, 2019.

[3] Abdelmoula Dmouj. Stock price modelling:theory and practice. *Vrije Universiteit Faculty of sciences Amsterdam*, 2006.

[4] P. Poncet et R. Portrait. *Finance de marché. 4e éd.* Broché, 2014.

[5] James D. Hamilton. *Time series Analysis*. Princeton, New Jersey : Princeton university, 1994.

[6] Neil Liberman. Decision trees and random forests. *Towards Data Science*, 2017.

[7] Marc Parizeau. *RESEAUX DE NEURONES (p. 8-13)*. PhD thesis, Université Laval, 2004.

[8] Eric Stellwagen Len Tashman. *ARIMA: The Models of Box and Jenkins*. 2013.