UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Prédiction de la surprise basée des valeurs fondamentales

dans le cadre d'une stratégie financière active

Auteure : Noémie Jamar
Promoteur: Corentin Vande Kerckove
Année académique 2021-2022
Travail de fin d'études (TFE) en vue d'obtenir le titre de
Master (60) en Sciences de Gestion
Horaire de jour

Remerciements

Je tiens à remercier Pr. Vande Kerckhove pour son soutien constant et ses remarques constructives. Un grand merci pour le temps qu'il a accepté de consacrer à nos rencontres régulières et pour les orientations de ce projet qu'il m'a encouragé à formaliser et à explorer.

Je suis reconnaissante envers tous les professeurs de l'UCLouvain pour tous les supports pédagogiques qui m'ont permis d'appréhender les différents domaines à analyser.

J'adresse également mes remerciements à ma famille pour son soutien permanent.

Table des matières

1 Introduction

Malgré les crises mondiales que nous traversons, que ce soit la crise COVID depuis début 2020 et puis la guerre en Ukraine en 2022, les marchés financiers restent un pilier du fonctionnement de notre société actuelle. En assurant les échanges entre les demandes d'investissement des entrepreneurs et les investisseurs sur base d'une maîtrise des risques, ils assurent le dynamisme de l'économie et permettent les évolutions dont notre monde a besoin. Le partage des informations sur le prix et la valeur des actions est une condition pour assurer le bon fonctionnement d'un marché financier dans un environnement concurrentiel et sain. C'est pourquoi les entreprises doivent communiquer régulièrement leurs résultats. Cela permet aux investisseurs de comprendre, sur base des valeurs fondamentales publiées, la situation, les opportunités et les risques. Ils peuvent ainsi différencier la valeur intrinsèque et la valeur marché de l'action. Cette dernière suit la loi de l'offre et de la demande. Afin d'aider les investisseurs dans leurs décisions, des analystes opèrent sur les marchés en identifiant le différentiel entre la valeur intrinsèque et le prix fixé par le marché. Ils donnent de manière régulière des estimations anticipées sur les résultats des entreprises comme les revenus, les résultats d'exploitation, ... Leurs différents avis sont pris en compte sous forme de moyenne et mis à disposition des investisseurs. L'établissement de ces valeurs anticipées moyennes se fonde, entre autres, sur les informations disponibles fournies par les entreprises, sur les résultats précédents, ou sur la tendance du secteur. Néanmoins, un effet d'émulation peut jouer entre analystes et établir un consensus qui peut être sur ou sous-estimée par rapport à la réalité.

1.1 Valorisation d'une action

La valorisation d'une entreprise ou de son action, réalisée par les analystes est primordiale pour assurer la transparence du marché et pour permettre la prise de décisions d'investissement. Cette valorisation peut se faire sur le principe du prix unique qui postule que sur un marché concurrentiel, deux actifs identiques vont tendre à avoir le même prix, encore faut-il identifier une entreprise de même taille, dans le même secteur, qui dispose des mêmes opportunités de croissance, du même endettement et fonctionne dans un même niveau de risque global. Une autre technique vise à identifier les flux financiers futurs selon le théorème du capital-valeur de Fisher qui stipule que le prix d'un actif financier doit être égal à la valeur actuelle des flux futurs auxquels a droit son propriétaire.La difficulté réside à identifier ces flux futurs. Dans cet ordre d'idée, le modèle Gordon-Shapiro s'appuie sur un taux de croissance des dividendes constant pour identifier les flux futurs et en déduire la valeur de l'action. Cependant, cette hypothèse n'est vérifiée que pour les entreprises qui ont une certaine stabilité de fonctionnement. Typiquement, cette hypothèse ne se vérifie pas pour les jeunes entreprises, elles paient rarement des dividendes les premières années, ensuite il faut plusieurs exercices comptables pour stabiliser

la croissance de leurs dividendes

Par ailleurs, la théorie de l'efficience des marchés introduite par Sharp, induit que le prix de l'action reflète toute la valeur de l'actif sur base des informations connues par le marché, que tout différentiel entre valeur et prix induit un arbitrage qui va annuler ce différentiel et ainsi réduire fortement la probabilité de trouver des différentiels. La conséquence de cette théorie est que la prédiction de valeur d'action ne peut pas se fonder sur de l'information nouvelle et devrait donc avoir les mêmes résultats qu'un choix aléatoire.

Beaucoup d'études ont pourtant montré par des exemples empiriques que cette hypothèse ne se vérifiait pas et qu'il était possible en utilisant des techniques statistiques ou de l'intelligence artificielle d'obtenir un meilleur résultat que celui prédit sur base simplement aléatoire. Selon [10], un grand nombre de données économiques fondamentales peuvent être prises comme prédictions significatives de la valeur de l'action. Alberg et al [1] déterminent que la prédiction sur une période de cinq ans via des réseaux neuronaux profonds, d'indicateurs fondamentaux de la santé d'entreprise que sont la valeur comptable normalisée par la capitalisation boursière et le résultat d'exploitation normalisé par la valeur de l'entreprise permet d'établir une stratégie de gestion de portefeuille plus performante que le marché.

1.2 Identification des différentiels de valeurs

Selon Wahlen et Wieland [9], des rendements supérieurs à ceux obtenus en suivant les recommandations consensuelles des analystes, peuvent être obtenus en analysant les publications des données fondamentales des entreprises et en prédisant grâce à ces données l'augmentation du bénéfice.

Ces techniques d'apprentissage ont été appliquées à la valeur de l'action afin de prédire un comportement à partir des leçons du passé sur une base assez large d'exemples. Cependant, la dynamique du prix de l'action est très fluctuante, non-linéaire avec à court terme beaucoup de mouvements perturbateurs liés à des facteurs externes qu'on peut qualifier de bruit. Selon [5], les données financières telles que la valeur d'une action démontrent des relations complexes, non-linéaires et sujettes à un rapport signal/bruit faible.

C'est pourquoi l'apprentissage sur base d'indicateurs économiques fondamentaux publiés de manière régulière, plutôt que de la valeur de l'action est une hypothèse qui devrait nous permettre une estimation plus précise de la dynamique d'évolution.

Selon Piotroski [8], des stratégies utilisant les publications des informations fondamentales des entreprises identifiant, ex ante, des différentiels entre les estimés des experts et la valeur finale publiée, permettent de sélectionner les

entreprises gagnantes et perdantes par rapport aux prévisions des experts et ainsi d'optimiser le rendement du portefeuille d'actions. Ces différentiels sont appelés surprises.

Par exemple, selon Becque [3], il est possible de prédire les surprises sur base d'une sélection d'entreprises pharmaceutiques, en utilisant le nombre d'estimés d'experts, le ratio max/mean et min/mean des prévisions d'analystes.

1.3 Recherche de la surprise

Ce travail s'inscrit dans la perspective d'identifier des entreprises qui lors de la publication de leurs chiffres, s'avéreront soit sous-évaluées par les marchés, soit sur-évaluées. On parle de sous-évaluation lorsque le prix marché est inférieur à la valeur intrinsèque et de sur-évaluation dans le cas contraire. Cette étude vise, à travers un apprentissage, à comparer l'usage de plusieurs indicateurs économiques fondamentaux prévus par les analystes et ceux publiés par les entreprises, afin de prédire un différentiel entre la valeur de revenu moyenne estimée par les analystes et la valeur réelle finalement publiée par l'entreprise. Ce différentiel est appelé surprise et sert d'indicateur aux sous-évaluations ou sur-évaluations d'actions qui procurent des opportunités de performance meilleure que la moyenne du marché.

L'objectif est donc de déterminer la meilleure méthode de prédiction de la surprise sur le revenu avant la publication des résultats finaux par l'entreprise. En utilisant une technique de prédiction via régression simple et apprentissage machine, nous allons en première approche, prédire directement la surprise de revenu et dans une approche alternative, prédire les indicateurs fondamentaux qui nous permettront ensuite d'établir une déduction de la surprise. Pour rappel, la surprise correspond au différentiel potentiel entre la valeur prédite par les analystes (le prix de l'action avant l'annonce des résultats de l'entreprise intègre déjà cette information connue par les marchés) et la valeur réelle annoncée (nouvelle information susceptible de produire une opportunité d'arbitrage sur le marché).

Le choix d'estimer cette valeur de surprise est motivé par plusieurs études montrant que cette variable a une importance significative pour anticiper l'évolution du cours de l'action. Selon Brown et al. [4], l'importance de la surprise n'a pas seulement un effet ponctuel de réajustement du prix de l'action sur le marché. Une surprise positive a aussi un effet sur le comportement des acteurs du marché qui augmentent leur intérêt par rapport aux informations de l'entreprise concernée de manière durable sur les périodes suivantes.

Selon Alvarado et al. [2], les surprises en matière de revenus et de bénéfices sont toutes deux importantes pour déterminer la performance des actions. Les bénéfices ont probablement plus d'importance pour une entreprise établie avec

une trésorerie saine, alors que la croissance du chiffre d'affaires d'une année sur l'autre est probablement plus importante pour une entreprise relativement jeune et en croissance. Le secteur, l'industrie, l'exposition au risque et la vision de l'équipe de direction sont évidement d'autres facteurs à considérer.

De plus, selon Ertimur et al. [6], l'information contenue dans la donnée *revenu* est plus pertinente que celle contenue dans la donnée des dépenses des entreprises. En effet, les dépenses, alimentées par plusieurs charges, sont plus fluctuantes et manipulables par les entreprises elles-mêmes. De plus, les investisseurs sont plus réactifs à une surprise sur les ventes plutôt que sur les dépenses, ce qui induit une plus forte réaction sur la valeur de l'action.

De même, selon Jegadeesh et Livnat [7], l'impact d'une surprise sur les revenus est plus marqué et plus persistant que celui d'une surprise sur les dépenses. Ce qui montre l'importance de la surprise sur revenu et notamment celle des années précédentes, pour l'analyse des investissements.

1.4 Question de recherche et méthode

Cette étude vise à prédire de manière la plus fiable possible les surprises en revenu des entreprises en utilisant les prévisions moyennes disponibles des experts. Plus précisément, la question de recherche de cette étude vise à déterminer si la prédiction de valeurs fondamentales comme le revenu ou la croissance de revenu permettant d'établir la surprise du revenu, est une méthode plus intéressante que la prédiction de la surprise du revenu elle-même. Le principe retenu est d'établir par apprentissage sur des résultats passés, des modèles d'évolution de cette surprise et de la valoriser chaque jour afin d'établir une liste pertinente d'entreprises à vendre ou acheter sur le marché des actions. L'idée est d'établir plusieurs modèles d'estimation de surprise en revenu et de comparer leur pertinence sur un jeu de données de test n'ayant pas servi à l'élaboration du modèle. Trois modèles vont être comparés : la prédiction de la surprise, l'établissement de la surprise sur base de la prédiction du revenu et l'établissement de la surprise sur base de la prédiction de la croissance annuelle du revenu. Afin de disposer d'une mesure de référence, la surprise établie l'année précédente pour le même trimestre, servira de valeur de comparaison. Pour vérifier la pertinence de ces modèles, plusieurs jeux d'apprentissage sur un nombre relativement important d'entreprises sont nécessaires.

Dés lors, un premier effort s'est porté sur la construction d'un journal contenant des données d'apprentissage. Le chapitre 2 présente les données qui ont été utilisées pour cette étude à partir d'une base de référence des informations sur les entreprises, Refinitiv. Il poursuit avec la classification des données qui a été utilisée pour éviter les *forward leak* càd l'utilisation, lors de l'apprentissage, d'informations qui apparaissent dans le futur de la période d'apprentissage et fausse la pertinence du modèle de prédiction. Ce chapitre présente également la

formule qui a été utilisée pour calculer la surprise et sa pertinence par rapport aux surprises publiées dans le référentiel de données financières Refinitiv. Il reprend les techniques utilisées pour l'élaboration du journal et le découpage mis en place pour obtenir trois jeux de tests d'apprentissage. Les séquences du traitement réalisé en python sont présentés en annexe.

Le deuxième volet de ce travail, expliqué dans le chapitre 3, a été consacré à la prédiction en présentant les techniques, les variables cibles, les variables caractéristiques et les métriques utilisées. Le chapitre 4 présente, pour un jeu d'apprentissage, les résultats de l'estimation des surprises, des revenus ou des croissances de revenu. Les résultats des autres jeux sont repris en annexe.

Finalement, sont présentées au chapitre 5, les comparaisons entre la surprise calculée à partir des données de test et les estimateurs de surprise : les trois surprises établies selon les trois modèles auxquelles on ajoute la surprise du trimestre de l'année précédente. La conclusion résume les résultats obtenus, les difficultés rencontrées et propose de nouvelles pistes à analyser.

2 Préparation des données

Afin de réaliser un apprentissage automatique, une première étape clé est la construction d'un jeu de données historiques fiables. Ce chapitre présente les données sélectionnées et les différentes transformations ou associations appliquées aux données extraites. Le risque de cette étape impose beaucoup de vigilance afin de respecter le timing d'apparition des informations. Ainsi les données associés sont classifiées en tant que données caractéristiques (connues au moment de la prédiction) et données cibles (appartenant au futur). En effet, les informations apparaissent sur le marché à des moments précis. Il est donc important dans la préparation du jeu de données d'apprentissage de bien respecter le timing d'apparition de ces informations. Si une donnée du futur transparaît dans le jeu d'apprentissage, cela fausse complètement le processus de recherche des dynamiques. Il est probable que la prédiction donnera un très bon résultat mais la prédiction de la dynamique ne fonctionnera pas pour un autre jeu de données ou de situation.

L'outil source des informations est bien connu des financiers : Refinitiv est un des plus grands fournisseurs de données sur les marchés financiers. Sa plateforme permet l'accès aux informations disponibles sur les marchés et les publications des entreprises.

2.1 Extraction de Refinitiv

Les données extraites de Refinitiv portent sur la période allant de janvier 2017 à la date de l'extraction réalisée en mars 2022. Le choix de ces dates est arbitraire. Il est important de pouvoir travailler sur une période temporelle de quelques années pour obtenir par apprentissage des tendances d'évolution. Par contre, il faut aussi tenir compte des temps de traitement qui augmentent avec le volume des données. Les extraits concernent les 500 entreprises S&P500 cotées aux États-Unis. Refinitiv permet l'extraction des informations soit via des requêtes manuelles soit via un lien qui capte les informations en continu. Dans le cadre de cette étude, des requêtes ont été utilisées afin de produire des fichiers excel qui ont été mis en forme à l'aide du programme python.

Les données extraites se retrouvent sous la forme de liste de records (ou lignes) identifiés par une clé unique constituée de la compagnie *cpy* et la *Period-End-Date*, c'est à dire la date de fin du trimestre concerné.

Données	**Couverture**	**Période**	**Fréquence**
Identifiant entreprise (S&P500)	/	/	unique
Revenu publié	Quarter	2017-2022	1/trimestre
Prévision moyenne du Revenu	Quarter	2017-2022	journalière
Surprise du revenu	Quarter	2017-2022	1/trimestre
EBIT	Quarter	2017-2022	1/trimestre
Prévision moyenne EBIT	Year	2017-2022	journalière
Classification nombre d'employés et secteur d'activité	/	04-2022	unique
Résultat annuel ESG	Year	2017-2022	1/an

TABLE 1 – Ensemble des données extraites de Refinitiv

La table 1 présente la liste des données extraites. On trouve ainsi, entre autres, les identifiants des entreprises concernées par l'analyse, les revenus publiés par trimestre et entreprise, les estimations moyennes de revenus émises par les experts en anticipation des résultats effectifs, les résultats d'exploitation EBIT publiés par entreprise et trimestre, les estimations moyennes de résultat d'exploitation (EBIT) des entreprises émises par les analystes. Chaque information publiée par une entreprise ou émise par les analystes est associée à une date qui correspond au moment de mise à disposition de l'information pour les marchés. Typiquement, le revenu publié (*Revenue-Actual*) est associé à la date de publication laquelle sera renommée en *FilingDate* dans le traitement.

De plus, des données considérées comme quasi-statiques ont été extraites afin de qualifier l'entreprise : nombre d'employés, secteur d'activité, reporting annuel ESG. Le nombre d'employés est utilisé pour qualifier la taille de l'entreprise en terme de ressources humaines. Le secteur d'activité est utilisé pour faire des regroupements et analyser de manière plus détaillée si des logiques peuvent s'attacher à des secteurs plutôt qu'à l'entreprise. Le reporting ESG est un indicateur de gouvernance et de transparence de l'entreprise au delà des informations financières qui vise à démontrer l'incarnation de la responsabilité sociétale dans le chef de l'entreprise. Les deux premières données ont été associées à l'entreprise sous forme constante pendant la période d'analyse, le reporting ESG a été associé à l'année auquel il se rapporte.

Nous utiliserons les surprises [1] en revenu établies par Refinitiv par entreprise et trimestre pour contrôler la méthode de calcul de la surprise choisie. En effet, Refinitiv publie la surprise en revenu à la date de publication des résultats

1. La surprise est la différence entre l'information de revenu publiée par l'entreprise et la dernière estimation de revenu établie par les experts.

Celle-ci est établie en comparant le revenu publié au dernier estimé moyen des analystes. Afin de disposer d'un calcul de surprise pour chaque jour de publication d'estimé moyen des analystes, nous avons utilisé une formule permettant d' établir nous-même la surprise. Nous comparerons nos résultats aux surprises extraites de Refinitiv afin de contrôler la pertinence de notre formule.

2.2 Établissement du journal

Dans les données extraites, certaines ont une fréquence d'apparition trimestrielle, ce sont les données réelles ou *actual* en anglais, publiées par les entreprises, d'autres comme les estimés moyens des analystes sont établies quasi-journalièrement. Pour rappel, l'objectif de la préparation des données est principalement d'établir un journal des informations connues à la date du jour, appelées variables caractéristiques (features) tout en les reliant aux informations finales du trimestre concerné, connues à postériori (à la date de la publication des résultats de l'entreprise) lesquelles serviront comme objectifs d'apprentissage ou cibles (target).

Ce journal est construit en lignes dont la clé d'identification est l'association de la date du jour du journal, de l'entreprise et du trimestre concerné. Ces éléments sont référencés comme suit : (*Date*,*cpy*,*Period-End-Date*).

Afin d'illustrer la construction du journal, nous utilisons l'exemple de l'entreprise A pour le trimestre juillet 2020. La figure 1 illustre les données pour l'entreprise A, du trimestre en question, le trimestre précédent ainsi que celui de l'année précédente. Pour construire le journal, la technique de *merge* appelée aussi join, permet de compléter une ligne en reliant les informations provenant d'autres fichiers en respectant une clé de rapprochement définie. La figure 2 montre l'ajout des informations finales par jointure avec les clés (*cpy*, *Period-End-Date*), (*cpy*, *PreviousQuarter*) ou (*cpy*, *PreviousSeasonalQuarter*).

Revenu Publié

cpy	Period End Date	FilingDate	Revenue - Actual
A	31/07/19	14/08/19 16:05	1274000000
A	30/04/20	21/05/20 16:08	1238000000
A	31/07/20	18/08/20 16:05	1261000000

Prévision moyenne du Revenu

cpy	Period End Date	Date	Revenue - Mean
A	31/07/20	21/05/20	1251306920

Code couleur

Information établie à postériori lors de la publications des résultats du trimestre de l'entreprise

Information connue à la date du journal

FIGURE 1 – Données pour l'entreprise A et le trimestre juillet 2020

Etape 1 : Regroupement des informations liées au trimestre d'une entreprise

---------clé---------

Calculé à partir de Period End Date

cpy	Period End Date	FilingDate	Revenue – Actual	EBIT-Actual	PreviousSeason Quarter	Previous Quarter
A	31/07/20	18/08/20 16:05	1261000000	298000000	2019-07-31 00:00:00	2020-04-30 00:00:00

Join avec cpy+Previous SeasonQuarter | Join avec cpy+Previous Quarter | Join avec cpy+Previous SeasonQuarter | Calculé | Calculé

previous season Revenue Actual	previous Revenue Actual	previous season EBIT	Growth Actual	growthSeasonality Actual
1274000000	1238000000	280000000	0,018578352	-0,010204082

FIGURE 2 – ETAPE 1 : intégration des données cibles du cas exemple

Etape 2 : **Journal (préparation d'un extrait pour un trimestre d'une entreprise)**

Date	Cpy	Period End Date	FilingDate	...
Previous Quarter FilingDate= 21/05/20	A	31/07/20	18/08/20 16:05	
...	A	31/07/20	18/08/20 16:05	
...	A	31/07/20	18/08/20 16:05	
...	A	31/07/20	18/08/20 16:05	
...	A	31/07/20	18/08/20 16:05	
...	A	31/07/20	18/08/20 16:05	
FilingDate= 18/08/20	A	31/07/20	18/08/20 16:05	

FIGURE 3 – ETAPE 2 : préparation du journal

Afin d'établir le journal comme présenté dans la figure 3, il est nécessaire de préciser pour le trimestre d'une entreprise, une plage de jours possibles pour l'émission d'estimations : le premier jour correspond à la date de publication des résultats du trimestre précédent et le dernier jour à la date de publication des résultats de ce trimestre.
Pour ce faire, par entreprise et trimestre, un extrait du journal établit une ligne pour chaque date (jour ouvrable) entre la date de publication du trimestre précédent jusqu'à la date de publication du trimestre concerné. Ensuite, il est possible d'associer les valeurs estimées à une certaine date par les analystes par une jointure via la clé (*Date*,*cpy*,*Period-End-Date*) comme présenté dans la figure 4. Comme toutes les dates n'ont pas reçu d'estimés, la technique de *forwardfill* permet de remplir les variables "estimés" vides en reprenant selon le sens chronologique le dernier estimé émis précédemment jusqu'à ce qu'un nouvel estimé apparaisse dans le journal. Ensuite,tous les sous-journaux construits pour chacun des trimestres d'une entreprise sont concaténés pour établir le journal complet.

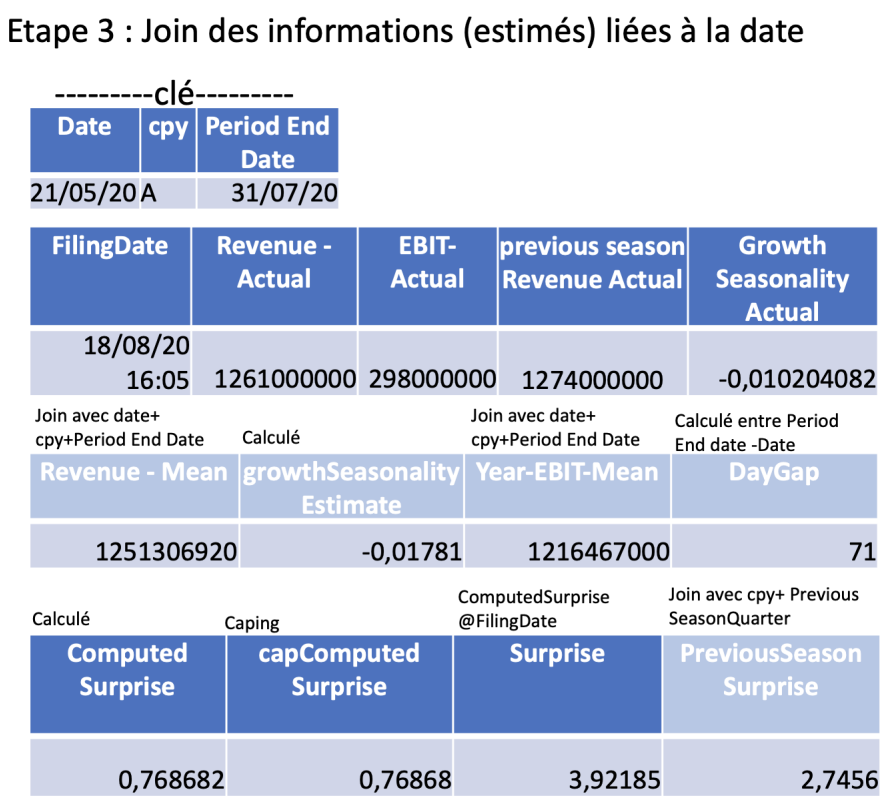

FIGURE 4 – ETAPE 3 :intégration des données caractéristiques du cas exemple

Deux types de données sont à différencier dans ce journal afin de distinguer les données connues à la date du journal (données journalières) et les données qui seront connues lors de la publication des résultats du trimestre de l'entreprise (données finales). On parle de variables caractéristiques (*features*) pour les données journalières et de cible (*target*) pour les données finales. Sans cette précaution, l'apprentissage risque d'être faussé par des informations futures (non connues à la date du journal) dans le jeu des données d'apprentissage.

Les données qui servent de cible sont le revenu publié par l'entreprise (*Revenu Actual*) ou l'EBIT publié par l'entreprise (*EBIT Actual*) pour le trimestre en question. Ces données sont liées à la clé *cpy+Period-End-Date*. A la date du journal (*Date*), elles ne sont pas connues mais serviront de cible dans le cadre de l'apprentissage du modèle.

Les données caractéristiques (features) connues à la date du journal pour le trimestre de l'entreprise concerné sont les estimés de revenu ou d'EBIT établis à la date du jour ou précédemment, les publications de revenu et d'EBIT du trimestre précédent et du trimestre de l'année précédente (Y2Y en abrégé) ou les calculs effectués à partir de ces données.

Afin de faciliter le processus d'apprentissage, une variable supplémentaire a été calculée pour indiquer la différence entre la date courante du journal et la date de fin du trimestre (catégorie : données journalières caractéristiques, nom : *DayGap*).

2.3 Calcul des croissances

Sur base des données de ce journal, plusieurs éléments ont été calculés selon les formules présentées ci-dessous. En utilisant l'information de revenu publié à la fin du trimestre, il est possible de calculer deux autres variables cibles : la croissance du revenu par rapport à celui de la période précédente (catégorie : données finales ou target, nom : *growthActual*) et la croissance du revenu par rapport à celui du trimestre identique de l'année précédente, Y2Y en abrégé, (catégorie : données finales ou target, nom : *growthSeasonalityActual*).

$$growthActual = \frac{\text{Revenu Actual} - \text{Previous Quarter Revenue actual}}{\text{Previous Quarter Revenue actual}} \tag{1}$$

$$gthSeasonActual = \frac{\text{Revenu Actual} - \text{Previous Season Quarter Revenue actual}}{\text{Previous Season Quarter Revenue actual}} \tag{2}$$

Deux nouvelles variables caractéristiques ont été calculées sur base de données connues à la date du journal : la croissance du revenu estimé par rapport à celui publié de la période précédente (catégorie : données journalières ca-

ractéristiques, nom : *growthEstimate*) et la croissance du revenu estimé par rapport à celui publié le trimestre correspondant de l'année précédente (catégorie : données journalières caractéristiques, nom : *growthSeasonalityEstimate*). Ce sont les croissances sur base annuelle (Y2Y), qui ont été préférées dans nos analyses pour tenir compte des particularités des entreprises qui démontrent des différences importantes selon la saison.

$$growthEstimate = \frac{\text{Revenu Mean} - \text{Previous Quarter Revenue actual}}{\text{Previous Quarter Revenue actual}} \tag{3}$$

$$gthSeasonEstimate = \frac{\text{Revenu Mean} - \text{Prev Season Quarter Revenue actual}}{\text{Prev Season Quarter Revenue actual}} \tag{4}$$

2.4 Calcul des surprises

Notre objectif est d'identifier jour par jour, avant publication, les entreprises qui publieront un revenu fort éloigné de la prédiction des analystes, créant ainsi une surprise et donc, in fine, une modification de la valeur de l'action. Refinitiv publie une valeur de surprise uniquement disponible à la date de publication. Cependant, pour tenir compte des fluctuations éventuelles des estimés des experts, nous avons besoin de prédire la surprise tous les jours du journal sur base de notre jeu de valeurs.

A partir des données publiées et l'estimé de revenu, la surprise calculée chaque jour du journal porte le nom de *computedSurprise* (catégorie : données finales cibles).

$$ComputedSurprise = 100 * \frac{\text{Revenu Actual} - \text{Revenu estimé}}{\text{Revenu Actual}} \tag{5}$$

D'autres formules auraient été possibles comme le rapport du revenu estimé moyen sur le revenu publié finalement. Nous avons réalisé plusieurs essais pour construire une formule et avons retenu la formule permettant de produire une surprise qui s'approche de celle publiée par Refinitiv. A partir de ComputedSurprise, trois nouvelles variables ont pu être établies. La première variable utilise la même formule mais appliquée le jour de la publication des résultats et indique ainsi la surprise par rapport au dernier estimé des analystes avant publication par l'entreprise (catégorie : données finales cibles, nom : *Surprise*).

La seconde variable correspond en majorité aux valeurs de *computedSurprise*, seules les 0.005 pourcent de valeurs extrêmes ont été atténuées à la dernière valeur admissible (catégorie : données finales cibles, nom : *capcomputedSurprise*) par la technique de caping. La troisième variable correspond

à la surprise de la période Y2Y précédente (catégorie : données journalières caractéristiques : nom : *Previous seasonal surprise*). Cette dernière variable est connue à la date du journal.

Voici le graphique XY, figure 5, reprenant les données de surprise selon Refinitiv ainsi que celles établies par notre formule et prises à la date de parution des données de revenus réels.

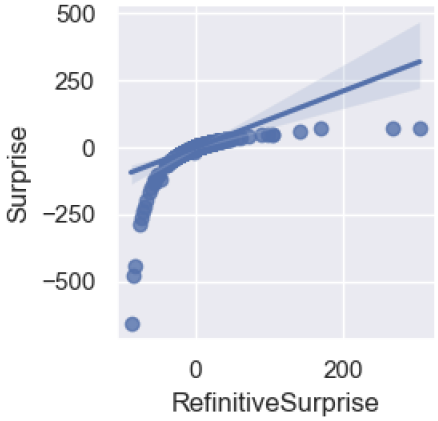

FIGURE 5 – corrélation entre la surprise Refinitiv et celle calculée, établie au jour de publication

En reliant pour chaque trimestre par entreprise, les deux valeurs, celle produite par Refinitiv et celle calculée, nous avons obtenu un facteur de corrélation via le facteur r de Pearson de 0.64398, ce qu'il indique une lien entre les variables mais pas une forte corrélation laquelle serait caractérisée par un facteur de 1. Quand on observe le graphe de la figure 5, on constate que la relation entre les deux surprises n'est pas linéaire.

Après analyse des graphiques sur la figure 5, on observe que les différences notables sont surtout situées dans les valeurs extrêmes. C'est pourquoi nous avons entrepris de *caper*, bloquer les valeurs extrêmes sur un certain pourcentage de ces valeurs extrêmes. La meilleure adéquation est obtenue avec un blocage des 0.005 pourcent de valeurs les plus grandes et les plus petites. Dans ce cas, le coefficient de corrélation selon le r de Pearson s'établit à 0.79475.

Pour valider notre formule, nous avons également vérifié le facteur de Spearman qui indique 0.99906 (p-value $< 1\epsilon^{-16}$) avec la surprise capée. Ce facteur indique une relation monotone entre les rangs des deux variables. Cela démontre que la formule que nous avons retenu est très proche de celle utilisée par Refinitiv.

La valeur de surprise calculée selon la formule présentée et capée à 0.05 pourcent est utilisée pour l'apprentissage dans le fichier train afin d'établir la cible. Par contre, c'est la surprise calculée (computedSurprise) qui est utilisée pour établir les évaluations de la pertinence des modèles sur le jeu de test.

2.5 Contrôle de qualité

Pour assurer les contraintes nécessaires afin d'effectuer les calculs (dénominateur) et garantir un travail de prédiction fiable, il est important de s'assurer de ne pas disposer de données nulles ou vides.

Données	**Taille initiale**
Entreprises (S&P500)	504
Revenu publié	10.023
Prévision moyenne Revenu	130.492
Surprise du revenu	10.023
EBIT	9.987
Prévision moyenne EBIT	122.491
Classification nombre d'employés et secteur entreprise	503
Résultat annuel ESG	2.506

Table 2 – Volumétrie des données extraites de Refinitiv

La table 2 présente la taille initiale des données extraites de Réfinitiv, c'est à dire, le nombre de lignes présentes dans le fichier. Il n'est pas surprenant que le nombre d'estimés soit plus important que le nombre de publications. En effet, des estimés peuvent être émis pendant les 90 jours et parfois plus, du trimestre concerné alors que la publication se fait une seule fois par trimestre.

Données	**Taille après nettoyage**
Revenu publié	9.605
Prévision moyenne Revenu	127.820
Surprise du revenu	10.023
EBIT	9.981

Table 3 – Réduction de volume des données extraites de Refinitiv

La table 3, présente la réduction volumétrique des données des fichiers de revenu publié, de la prévision moyenne de revenu, de l'EBIT publié suite au nettoyage préliminaire permettant d'éliliner les données nulles ou vides.

L'établissement du journal, comme expliqué précédemment, augmente le nombre de ligne puisque il prend en compte non seulement les dates pour lesquels un estimé est émis par les spécialistes mais aussi d'autres dates, entre la date de publication du trimestre précédent et celle du trimestre concerné, en utilisant la technique de *forwardfill* permettant de tenir compte du dernier

estimé émis à cette date. Le fichier passe donc d'environ 120.000 émissions d'estimés à plus de 487.000 lignes dans le journal.

Lors de la construction du journal en incluant les informations sur la période précédente et sur la période équivalente de l'année précédente, puis dans un second temps en reliant les données à la surprise de l'année précédente Y2Y (*previous season surprise*), nous avons fait le choix d'éliminer les lignes qui n'ont pas reçu de valeurs. Ce choix de nettoyage à deux niveaux pour le calcul des croissances puis pour la surprise de l' année précédente provoque une forte réduction des volumes. On observe une réduction jusqu'à 347K lignes du journal. Dans le fichier intermédiaire reprenant uniquement les surprises calculées à la date de publication des revenus et servant à la comparaison avec les surprises de Refinitiv, le nombre se réduit à 5.393 à comparer aux 10.000 trimestres initialement pris en compte.

	Taille initiale	**Retrait des valeurs Y2Y vides**
Taille du journal des estimés	487.880	347.290
Taille du journal des surprises	7.375	5.393

TABLE 4 – Réduction de la volumétrie suite aux traitements de contrôle

Ces différents contrôles ou nettoyages en cascade ont eu pour effet non seulement de réduire la volumétrie des données comme présenté dans la table 4, mais aussi de limiter la couverture temporelle des données présentes dans les fichiers comme montré dans la table 5.

Les données avant traitement couvrent les périodes de 2017 à 2022. Les traitements de cohérence réduisent le journal aux dates de octobre 2018 à 2022 afin de disposer à la fois de la croissance du revenu lors du trimestre de l'année précédente et de la surprise de l'année précédente. Il serait bon de revoir le plan de la création du journal en reportant le plus possible les nettoyages afin de limiter cette réduction en évitant l'effet cascade.

Traitement	Date la plus ancienne	Date la plus ancienne après nettoyage	Remarque
étape 1 :join avec les revenus Y2Y : voir figure 1	2017	2018-04	réduction nécessaire pour disposer de Y2Y growth
étape 3 :Calcul de Computed Surprise : voir figure 4	2018-04	2018-10	réduction nécessaire pour disposer de previous season surprise
Résultat final : Journal	2018-10		

TABLE 5 – Evolution de la couverture temporelle des données suite aux regroupements et aux contrôles

2.6 Préparation à l'apprentissage

Dans le cadre de l'apprentissage, il est nécessaire de diviser le journal en deux fichiers, l'un qui a vocation à être utilisé pour l'apprentissage (train) et l'autre permet de valider le modèle retenu (test).

fichier	nbr de lignes (date-cpy-trimestre)	nbr d'entreprises (cpy)	nbr moyen de dates par entreprise	nbr de dates (uniques) dans le journal	première date	dernière date	nbr d'association entreprise - trimestre = nombre de publications	nbr moyen de jours du journal par trimestre d'une entreprise
journal	347290	486	715	837	31/12/18	31/01/22	5393	64

FIGURE 6 – Données métriques du contenu du journal

La figure 6 présente pour le journal établi précédemment, le nombre de lignes correspondant à la clé de référence unique de chaque record du fichier (Date,Cpy=entreprise, Period End Date= trimestre concerné), le nombre d'entreprises présentes, le nombre moyen de lignes concernant une entreprise, le nombre de dates (uniques) du journal, la première date et la dernière date, le nombre d'association entreprise-trimestre et le nombre de jours avec des estimés par trimestre d'entreprise. Sur base de ce journal, un fichier train et un fichier test ont été préparés pour trois jeux d'apprentissage en sélectionnant une séparation au 31/12/19 et au 31/12/20. La volumétrie des fichiers respectifs est présentée à la figure 7.

La découpe de ce journal doit être réalisée en comparant la date de coupure,

à 2 dates dans les lignes du journal : la date courante appelée *Date* et la date de publication des données finales du trimestre appelée *FilingDate*. Le fichier train est établi en reprenant les lignes ou records pour lesquels ces 2 dates sont antérieures à la date choisie comme séparation, alors que le fichier test est constitué des lignes ou records dont les deux dates sont postérieures à la date de séparation choisie.

Cette coupure large comme présenté à la figure 8, permet de garantir qu'aucune ligne de train ne permet de faire découvrir à la méthode d'apprentissage un résultat cible qui ne devrait être que dans le test. Cela revient à valider que le dernier trimestre présent dans le fichier train, doit être antérieur au premier trimestre repris dans le fichier test, comme visible sur la figure 9. Cette précaution permet d'éviter les forward leak ou information du futur mise à la disposition du système d'apprentissage.

fichier	nbr de lignes (date-cpy-trimestre)	nbr d'entreprises (cpy)	nbr moyen de dates par entreprise	nbr de dates (uniques) dans le journal	première date	dernière date	nbr d'association entreprise - trimestre = nombre de publications	nbr moyen de jours du journal par trimestre d'une entreprise
train_19	89172	475	187,73	254	1/01/19	20/12/19	1390	64,15
test_20	91110	472	193,03	256	1/01/20	23/12/20	1428	63,8
train_20	101721	476	213,70	256	1/01/20	23/12/20	1842	55,22
test_21	88645	469	189,01	254	1/01/21	22/12/21	1397	63,45
train_19_20	211266	480	440,14	517	1/01/19	23/12/20	3232	65,37
test_21	88645	469	189,01	254	1/01/21	22/12/21	1397	63,45

FIGURE 7 – Données métriques du contenu des fichiers d'apprentissage

Date	cpy	Period End Date	FilingDate	Train Coupure=31/12/19	Test Coupure=31/12/19
09/10/19	A	30/09/19	10/10/19 16:04	OK	NOK
21/12/19	A	31/12/19	18/01/19 16:05	NOK	NOK
01/01/20	A	31/03/20	07/04/20 16:06	NOK	OK

FIGURE 8 – principe de préparation des fichiers train et test

fichier	dernier trimestre présent	premier trimestre présent
train_19	30/11/19	
test_20		31/01/20
train_20	30/11/20	
test_21		31/01/21
train_19_20	30/11/20	
test_21		31/01/21

FIGURE 9 – Contrôle du non recouvrement en trimestre des fichiers train et test

En général dans les méthodes d'apprentissage, il est préférable de normaliser les données en les ramenant à leur moyenne proportionnellement à leur dispersion matérialisée par l'écart-type. Cependant cette technique, pourtant utile pour amener chaque variable caractéristique à un domaine de valeur comparable aux autres, a pour effet d'enlever une partie de l'information à la donnée. Dans cette étude, nous avons choisi de ne pas redimensionner les données excepté lors de l'analyse de l'importance comparative des variables caractéristiques (features) qui a, elle, nécessité ce traitement d'échelle.

Une différence par rapport à d'autres problématiques régulièrement résolues par intelligence artificielle est qu'il n'est pas aisé, dans cette étude, d'utiliser la technique de ré-échantillonnage connue sous le nom de validation croisée (cross validation). Les données devraient être réparties aléatoirement en 3 ensembles pour l'apprentissage, la validation et le test. Mais comme il faut éviter de prendre en compte des informations du futur lors de l'apprentissage et que le contrôle avec deux dates n'est pas aisé à automatiser, il est donc préférable de définir des frontières temporelles entre les données des différents fichiers.

3 Modèles de prédiction de la surprise

3.1 Techniques de prédiction

Afin de pouvoir réaliser la prédiction d'une valeur, il est nécessaire d'établir un apprentissage en sélectionnant un modèle et une mesure d'erreur qui sera optimisée. Cet apprentissage se réalise sur un jeu de données train qui contient les données caractéristiques et cibles. Ensuite, sur base du résultat de l'apprentissage et avec les données caractéristiques du jeu de données test, il est possible de prédire les données cibles. Ces dernières sont comparées aux données cibles réellement établies en fin de trimestre afin de qualifier le pouvoir de prédiction du modèle établi. Ce chapitre présente les modèles de prédiction utilisés, les variables cibles et caractéristiques choisies et les métriques à optimiser.

3.1.1 Régression linéaire multivariée

Le premier modèle utilisé dans cette étude est le modèle linéaire multivariables. C'est un modèle simple à mettre en oeuvre qui suppose une relation linéaire entre des variables indépendantes et la cible sous la forme d'une équation $Y = a + b_1{*}X_1 + b_2{*}X_2 + \ldots + b_n{*}X_n$, avec a l'intercept, b_i les coefficients des différentes variables indépendantes X_i. On parle de régression aux moindres carrés car la fonction de perte à minimiser est la MSE (mean squared error). [2]

3.1.2 Gradient Boosting

Le second modèle utilisé est basé sur la méthode d'apprentissage par *gradient Boosting*. Il est utile pour déterminer l'importance relative des variables caractéristiques dans la prédiction de la variable cible et ainsi vérifier la pertinence de leur choix. Le gradient boosting construit, par étape, un modèle de prédiction sous la forme d'un ensemble de sous-modèles de prédiction. La méthode se base sur la construction d'arbres de décision qui structure les variables et leurs espaces de valeur afin d'établir une prédiction en optimisant la fonction de perte MSE : Mean Squared Error qui sera présentée au paragraphe 3.4. L'ensemble de plusieurs sous-modèles de prévision associés, permet de former un modèle plus fort qui minimise la fonction de perte plus que chaque sous-modèle individuellement. En effet, dans le cas du gradient boosting, un nouvel arbre de décision est construit à chaque étape pour minimiser les erreurs résiduelles de l'ensemble de l'étape précédente. Le taux d'apprentissage et le nombre d'itérations sont des paramètres de contrôle du processus d'optimisation. Le taux s'applique au nouvel arbre pour réaliser un apprentissage par petit pas et le gradient permet d'identifier si la direction prise en terme d'optimisation doit être prolongée. Cette méthode est assez longue car itérative mais elle obtient de bons résultats à condition qu'il existe des corrélations entre les variables caractéristiques. Dans notre cas d'étude, les données croissance de

2. Cette métrique est expliquée dans la section 3.4

revenu et revenu ne sont pas des variables totalement indépendantes, elles se rapportent à la même entreprise et sont liées par des formules de calcul. Ce modèle permet d'établir l'importance des variables caractéristiques dans la prédiction. Plus une variable caractéristique est utilisée pour prendre des décisions clés dans les arbres décisionnels, plus son importance relative est élevée. On peut alors comparer les variables entre elles par la mesure de la diminution moyenne d'impuretés (MDI) que la variable permet, c'est à dire la somme des splits (pour tous les arbres) ou cette variable intervient, proportionnés à sa force de découpage.

3.1.3 Risques à couvrir par les méthodes de prédiction

Les risques que présentent traditionnellement les modèles de prédiction sont l'influence des données extrêmes et le risque de sur-ajustement. En effet, pour le premier risque, les valeurs de données aberrantes (fortement déviantes par rapport au plus grand nombre des autres données) peuvent avoir un impact sensible sur le résultat. Pour éviter ce risque, nous proposons d'éliminer les valeurs aberrantes de notre jeu d'apprentissage, en les bloquant à une valeur limite. Cette opération s'appelle la winsorisation. Elle modifie le jeu des données de sorte que les valeurs les plus extrêmes en proportion alpha sont "écrasées" sur les quantiles de niveaux $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$.

Le risque de surajustement aux données d'apprentissage signifie que le résultat du modèle est tout-à-fait adapté à l'ensemble des données d'apprentissage mais le modèle ne fonctionnera pas du tout pour un autre ensemble d'événements. Pour palier ce risque, il est nécessaire de disposer d'un ensemble important de valeurs d'apprentissage. L'utilisation de la cross-validation, c'est à dire le découpage de l'ensemble d'apprentissage en de multiples sections servant soit d'apprentissage soit de validation, serait une technique complémentaire pour réduire ce risque. Cependant, avec un journal temporel de données avec deux dates à contrôler, il est difficile de maîtriser le fait que des informations du futur ne soient pas divulguées dans les sections d'apprentissages. La balance bénéfice/risque nous a poussé à ne pas l'utiliser. Par ailleurs, une autre technique intégrée dans la méthode de prédiction, appelée régularisation, permet d'imposer une contrainte pour favoriser les modèles simples au détriment des modèles complexes. Ceci permet d'améliorer la généralisation de la solution. Dans le cas de la régression linéaire, la régularisation favorise des coefficients faibles.

3.2 Choix des cibles de prédiction

L'objectif in fine est de déterminer de manière journalière les entreprises pour lesquelles il serait judicieux de prendre une position d'achat ou de vente d'actions sur base d'un indicateur discriminant : la surprise sur le revenu du quadrimestre. Différentes techniques pour établir cette surprise sont mises en place : soit la prédiction directe de la surprise elle-même, soit la prédiction du revenu ou de la croissance Y2Y de celui-ci. Le calcul de la surprise est

ensuite réalisé pour chacune de ces prédictions. Il est préférable de comparer les quadrimestres respectifs d'année en année car comme nous l'avons précisé précédemment, les revenus de quadrimestre de certaines entreprises sont affectés par la saisonalité.

Plusieurs variables liées à un quadrimestre d'une entreprise ont donc été sélectionnées comme cible de prédiction : le revenu final, la surprise par rapport à la prédiction en cours, la croissance du revenu d'une année à l'autre.

3.3 Choix des variables caractéristiques

Certaines données provenant de l'ensemble d'apprentissage sont proposées comme variables prédictives. Ces dernières sont utilisées afin d'établir les caractéristiques du modèle de prédiction de la variable cible qui peut être le revenu, la croissance Y2Y ou la surprise.

La première variable envisagée, *Revenue-Mean*, correspond à la prédiction moyenne des experts disponible à la date du journal. Comme vous pourrez le constater dans le chapitre suivant, cette variable dispose d'une forte corrélation avec la variable cible de revenu du quadrimestre considéré de l'entreprise.

La croissance de la prédiction moyenne des analystes par rapport au quadrimestre de l'année précédente, *SeasonalGrowthEstimate*, est une seconde variable possible. Elle correspond au calcul de croissance entre la variable *Revenue-Mean* et le revenu publié au quadrimestre identique de l'année précédente. Cette variable a été choisie pour donner au modèle un estimatif de croissance bien que la corrélation entre ces deux variables est très faible, comme signalé précédemment.

Deux autres variables sont la prédiction moyenne des analystes en terme de EBIT annuel, *Year-EBIT-Mean* et la surprise en revenu Y2Y, *PreviousSeasonSurprise*. L'EBIT est la différence entre revenu et charges pour l'entreprise dans un quadrimestre. Cette valeur fondamentale pour l'entreprise apporte une information qui induit le bénéfice qu'une entreprise tire de son revenu. La surprise Y2Y correspond à la surprise établie entre la dernière prédiction moyenne des experts et le revenu publié à la date de publication pour le quadrimestre identique de l'année précédente. Elle donne de l'information sur la surprise à laquelle on doit s'attendre.

La différence de temps entre la date de prédiction et celle de fin de quadrimestre est la variable *Daygap*. Cette dernière variable spécifie le nombre de jour entre la date du journal où l'on considère la prédiction moyenne des experts et la date de fin de quadrimestre. Elle permet au modèle de déterminer si les informations fournies sont proches ou loin de la fin de quadrimestre. Elle peut être positive si la date où la prédiction est active se situe pendant le trimestre concerné et négative si la date est postérieure à ce trimestre.

3.4 Métriques d'évaluation de la prédiction

En général, lors d'un apprentissage, le modèle de régression est optimisé de manière à minimiser une métrique d'évaluation. Dans le cas d'une régression linéaire, cette métrique correspond à la MSE (Mean Squared Error) ou erreur quadratique moyenne.

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (e_i)^2 = \frac{1}{n} * \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

avec n le nombre de date du fichier test et e_i l'erreur entre la valeur cible réelle y_i et la valeur cible prédite $\hat{y}_i$.

La métrique RMSE est la racine carrée de l'erreur quadratique moyenne. Cette métrique est très sensible aux données aberrantes car elle pénalise plus fortement les grandes erreurs. La RMSE est utilisée pour mesurer l'erreur de prédiction notamment lors des exercices d'apprentissage. Il s'agit de sélectionner le modèle dont la valeur de RMSE est le plus faible possible.

Le score R2, appelé coefficient de détermination, quantifie la performance d'un modèle de régression.

$$R2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{7}$$

On peut l'assimiler à l'erreur du modèle envisagé, divisé par l'erreur d'un autre modèle qui prédirait en permanence la moyenne de la variable à prédire. Cela permet de mesurer le rapport de la variance expliquée par le modèle, sur la variance totale. Plus le score R2 s'approche de 100 pourcent, plus le modèle est performant. Un modèle qui prédirait tout le temps la moyenne aurait un score de 0. Un R2 négatif signifie que les prédictions sont moins bonnes que si l'on prédisait systématiquement la valeur moyenne. Le score R2 est utilisé pour comparer plusieurs modèles entre eux.

3.5 Méthodes d'établissement de la surprise anticipée

Pour rappel, l'objectif de cette étude est de prévoir la surprise chaque jour afin de décider au mieux de la politique d'investissement en actions. Selon notre approche, cette surprise est estimée selon trois méthodes différentes. L'estimateur *predicted surprise* se construit par la prédiction de la surprise sur base des variables caractéristiques connues à la date du journal. Les estimateurs *Surprise by RA* et *Surprise by G* correspondent respectivement à la surprise établie par calcul après prédiction du revenu (*RA=Revenue Actual*) et à celle établie à partir de la prédiction de la croissance Y2Y de revenu (*G=seasonal growth*).

$$\text{Surprise by RA} = 100 * \frac{\text{Revenu Prédit} - \text{Revenu estimé}}{\text{Revenu Prédit}} \tag{8}$$

$$\text{RevDéduitG} = \text{PrevSeasonRevActual} + \text{SeasonGthPredicted} * \text{PrevSeasonRevActual} \tag{9}$$

$$\text{Surprise by G} = 100 * \frac{\text{Revenu déduit G} - \text{Revenu estimé}}{\text{Revenu déduit G}} \tag{10}$$

L'estimateur *previous season surprise* sert, quant à lui, de référence de comparaison.

Afin de mesurer la qualité en information de ces estimateurs, deux métriques *hit rate* et *anchored hit rate* sont analysées pour les trois jeux d'apprentissage et de test.

Ces trois estimateurs ainsi que la référence doivent permettre d'anticiper la surprise, ils seront comparés à la valeur de la variable *computedSurprise* qui correspond à la surprise de l'estimé des experts à cette date, établie à postériori lorsqu'on connaît le résultat final de l'exercice trimestriel de l'entreprise.

3.6 Métrique d'évaluation des méthodes d'établissement de la surprise

Pour comparer les méthodes d'établissement de surprise, nous allons établir une métrique en *hit rate* et ensuite en *anchored hit rate*. Le *hit* vérifie que la surprise anticipée par un de nos trois modèles, a le même signe que la surprise calculée à postériori à partir des valeurs finales du quadrimestre. Le taux de *hit* permet de mesurer sur les données du fichier test, le résultat en hit des différentes surprises : *predicted surprise*, *Surprise by RA*, *Surprise by G* et de comparer à la référence *previous season surprise*. Comme la finalité en calcul en surprise est d'identifier les actions d'entreprises à acheter, vendre ou conserver dans une approche dynamique de gestion du portefeuille, le fait que la proposition de surprise ait le même signe que la surprise réelle conforte son choix en tant que paramètre de décision.

La mesure en *anchored hit rate* est construite sur le même principe que le hit rate mais à partir des variables recentrées par rapport à leur médiane. L'intérêt de cette dernière opération est de négliger le biais de surprise majoritairement positive car comme on pourra le voir dans le chapitre suivant, les analystes ont tendance à prédire une valeur en dessous du revenu réel. C'est cette dernière métrique que nous avons prise en compte pour comparer nos différents modèles de prédiction de la surprise.

4 Résultats empiriques

Ce chapitre présente les résultats empiriques obtenus sur base du journal élaboré à partir des données de Refinitiv. Pour rappel, nous voulons in fine prédire la surprise en revenu. Pour ce faire, nous utilisons trois méthodes : prédire la surprise elle-même, prédire le revenu et en déduire la surprise ainsi que prédire la croissance en revenu pour en déduire également la surprise. Pour valider ces modèles, nous comparons ces trois valeurs de surprise anticipée *Predicted Surprise*, *Surprise by RA*, *Surprise by G*, plus une référence qui est la surprise obtenue au même trimestre l'année précédente *Previous Season Surprise* à la surprise qui a été calculée dans le journal *Computed Surprise*.

Nous allons tout d'abord examiner les statistiques descriptives des variables cibles choisies pour la prédiction ainsi que celles des variables caractéristiques utiles. Nous étudierons ensuite les résultats des prédictions en surprise, revenu et croissance en analysant également l'importance des variables caractéristiques pour réaliser ces prédictions. Finalement, nous comparerons la distribution statistique des trois surprises obtenues pour identifier globalement les tendances. Nous analyserons finalement les métriques en hit rate et anchored hit rate qui permettent de comparer la force de prédiction des différentes surprises établies.

4.1 Statistiques descriptives des variables cibles

Cette section décrit les paramètres statistiques de la distribution des différentes variables cibles : revenu, surprise (*Computed Surprise* par jour) et croissance en les comparant à la donnée caractéristique qui devrait être la plus proche.

4.1.1 Revenu

	Mean	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Revenue-Actual	6.46E09	1.29E10	3.1E06	1.14E09	2.49E09	5.37E09	1.52E11
Revenue Mean	6.33E09	1.26E10	1.47E06	1.12E09	2.47E09	5.33E09	1.52E11

TABLE 6 – Statistiques descriptives des variables revenue (réel et estimé moyen des experts)

Les statistiques de la variable revenu réel (*Revenue-Actual*) sont comparées à celles de la variable revenu-estimé moyen des analystes (*Revenue-Mean*) dans la table 6. On observe une distribution assez proche. Pour déterminer si les deux variables sont liées, le coefficient de Pearson indique 0.998 de manière significative et le coefficient de Spearman est de 0.997. Ces valeurs indiquent que les deux variables sont fortement liées. Ceci permet d'anticiper une capacité de prédiction du revenu à partir de l'estimation moyenne en revenu des experts.

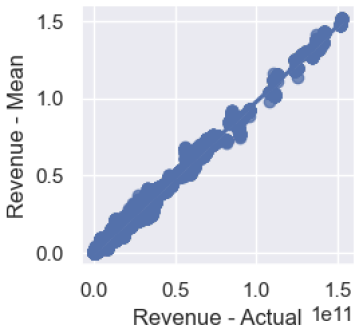

FIGURE 10 – Corrélation entre Revenu réel et Revenu estimé par les experts

La figure 10 montre qu'il existe un lien fortement linéaire entre les deux variables.

4.1.2 Surprise

Dans le journal de données nous disposons de deux variables Surprise : tout d'abord la surprise calculée à chaque date du journal *Computed Surprise*, elle compare le revenu estimé moyen des experts connu à cette date au revenu final (RA : *revenue Actual*) produit par l'entreprise, ensuite la *Surprise* prise à la date de la publication des résultats (*filingDate*). C'est cette dernière qui a été comparée à la surprise publiée par Refinitiv.

Par contre, pour le travail de prédiction, plutôt que la surprise finale du trimestre, nous prenons la valeur de surprise calculée journalièrement comme valeur cible d'apprentissage car l'objectif est de pouvoir prendre journalièrement une décision de vente ou achat d'action sur base des données estimées par les experts disponibles à ce jour.

La variable *Previous season Surprise* correspond à la variable *Surprise* mais au trimestre Y2Y précédent et nous sert de référence à l'anticipation de la surprise.

	Mean	**std**	**Min**	**25%**	**50%**	**75%**	**Max**
Daily Computed Surprise	0.39	23.07	-2359.88	-0.84	1.26	3.98	91.31
Previous Season Surprise	0.49	12.51	-475.26	-0.64	0.78	2.73	75.23

TABLE 7 – Statistiques descriptives des variables ComputedSurprise et Surprise de trimestre de l'année précédente

Le tableau 7 montre les statistiques de ComputedSurprise. On observe que plus de la moitié des surprises sont positives, ce qui signifie que les experts ont tendance à sous-estimer le revenu de l'entreprise. Ensuite, on observe que la corrélation avec l'estimateur PreviousSeasonSurprise est très faible tant en linéaire (Pearson : 0.01 p : 7,7E-10) qu'en terme de rangs (Spearman : 0,11). Une explication est que la variable computedSurprise contient de plus fortes variations observées pendant le déroulé du trimestre alors que PreviousSeason-Surprise est la photo finale.

FIGURE 11 – Corrélation entre ComputedSurprise et la Surprise établie sur le trimestre correspondant de l'an passé

Suite à la présence de données aberrantes, la figure 11 met en évidence qu'il n'y a pas de lien linéaire entre les deux variables.

Pour diminuer la déviance provoquée par ces dernières, nous utilisons la winsorisation de *computedsurprise* lors de l'établissement du modèle d'apprentissage. Cette technique ne s'applique pas sur les données de test pour respecter les conditions réelles de prise de décision financière.

4.1.3 Croissance de revenu sur une trimestre en comparaison à celui de l'année précédente

	Mean	std	Min	25%	50%	75%	Max
Y2Y Revenue growth	0.18	3.12	-0.99	-0.02	0.05	0.15	229.9
Y2Y Revenue growth Estimate	0.17	3.59	-1.00	-0.02	-0.04	0.12	286.77

TABLE 8 – Statistiques descriptives des variables seasonalgrowth (réel et calculé à partir de l'estimé moyen des experts)

La description des statistiques de la variable croissance de revenu réel (*seasonalGrowthActual*) est présentée dans le tableau 8 en parallèle avec celles de la variable croissance calculée journalièrement à partir de l'estimé moyen des analystes (*seasonalGrowthEstimate*).

Le coefficient de Pearson indique 0.986 de manière significative et le coefficient de Spearman est de 0.907. Ceci indique que les deux variables sont fortement liées.

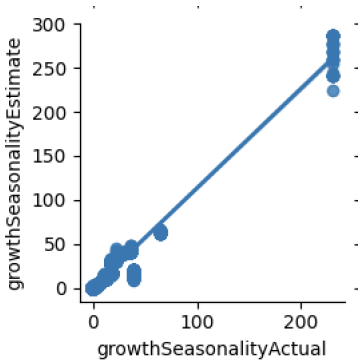

FIGURE 12 – Corrélation entre croissance de revenu réel et croissance de revenu calculée journalièrement à partir des estimés de revenu

Le graphique 12 suivant montre un lien linéaire entre les deux variables sauf dans le range des valeurs supérieures.

4.2 Statistiques descriptives des variables caractéristiques

	Mean	std	Min	25%	50%	75%	Max
Variables caractéristiques							
Revenue Mean	6.33E+09	1.26E+10	1.47E+06	1.12E+09	2.47E+09	5.33E+09	1.52E+11
DayGap	28.28	29.18	-114	5	33	54	102
Growth Seasonal Estimate	0.17	3.59	-1.00	-0.02	0.04	0.12	286.77
Previous Season Surprise	0.49	12.51	-475.26	-0.64	0.78	2.73	75.23
Year EBIT Mean	4.12E+09	7.83E+09	-1.05E+10	8.7E+08	1.74E+09	3.78E+09	1.11E+11
Variables connues et intégrées dans la prédiction des analystes							
Previous Season Revenue Actual	6.06E+09	1.18E+10	6.52E+06	1.06E+09	2.41E+09	5.22E+09	1.52E+11
Previous Season EBIT	9.71E+08	1.83E+09	-8.38E+09	2.07E+08	4.14E+08	9.34E+08	3.35E+10

TABLE 9 – Comparaison statistiques des variables caractéristiques potentielles

La table 9 montre les plages de valeurs des différentes variables caractéristiques à utiliser pour établir les prédictions. Les variables reprenant le revenu et l'EBIT publiés au trimestre identique de l'année précédente n'ont pas été retenues dans notre analyse car leur connaissance est déjà intégrée dans la variable *Revenue-Mean* des analystes financiers. On observe en calculant le maximum de la variable DayGap que la publication de résultats d'entreprise peut prendre plus d'un trimestre mais qu'en moyenne, cette publication se passe dans le mois suivant la fin du trimestre. On observe une médiane de croissance Y2Y du revenu de 4 pourcent avec une (ou des) valeur(s) maximum forte(s) qui pousse(nt) la moyenne vers le haut à 17 pourcent. Pour la surprise de la saison précédente, on observe de manière semblable que dans plus de la moitié des cas, la surprise est positive, ce qui signifie que les experts ont généralement tendance à sous-estimer le revenu de l'entreprise.

4.3 Résultats des prédictions pour chaque variable cible

Apprentissage Regression Linéaire	Cible	coefficient DayGap	coefficient Revenu - Mean	coefficient Season Growth Estimate	coefficient Year EBIT-Mean	coefficient Previous Season Surprise	Intercept
train_19_20,test_21	Revenu	-123425,70	1,01	103676433,00	0,01	8765202,54	-3,70E+07
train_19,test_20	Revenu	-188092,43	1,00	114065583,00	0,00	7663595,64	-1,61E+07
train_20,test_21	Revenu	239748,50	1,02	155233952,00	0,01	10825902,00	-5,64E+07
train_19_20,test_21	Computed Surprise	0,00	0,00	0,13	0,00	0,16	0,17
train_19,test_20	Computed Surprise	0,00	0,00	-0,16	0,00	0,13	-0,34
train_20,test_21	Computed Surprise	0,01	0,00	1,32	0,00	0,26	0,66
train_19_20,test_21	Y2Y Revenu Growth	0,00	0,00	1,00	0,00	0,00	0,00
train_19,test_20	Y2Y Revenu Growth	0,00	0,00	1,00	0,00	0,00	0,00
train_20,test_21	Y2Y Revenu Growth	0,00	0,00	1,02	0,00	0,00	0,01

FIGURE 13 – Coefficients et intercept des regressions

La figure 13 fournit les résultats des régressions linéaires sur les trois jeux de test sous forme de coefficients pondérateurs pour chacune des variables caractéristiques et d'un intercept correspondant à un biais appliqué à chaque prédiction.

On peut constater, pour la régression du revenu que toutes les variables caractéristiques ont des coefficients non nuls, alors que la droite de prédiction de la surprise utilise principalement la surprise du trimestre Y2Y précédent et l'estimation de croissance de revenu. Quant à la prédiction de la croissance de revenu Y2Y, elle s'appuie uniquement sur cette dernière variable caractéristique.

Apprentissage	**Cible**	**RMSE on test**	**R2 on test**
train 19-20, test 21	Revenu	1.16 E+09	0.99
train 19, test 20	Revenu	8.38 E+08	1.00
train 20, test 21	Revenu	1.39 E+09	0.9902
train 19-20, test 21	Computed Surprise	26.29	-0.02
train 19, test 20	Computed Surprise	34.16	0.00
train 20, test 21	Computed Surprise	28.13	-0.17
train 19-20, test 21	Y2Y Revenue growth	1.23	0.96
train 19, test 20	Y2Y Revenue growth	0.0911	0.87
train 20, test 21	Y2Y Revenue growth	1.33	0.95

TABLE 10 – Mesure de la prédiction (régression linéaire) sur les données test

L'analyse des métriques de l'erreur de prédiction, et du facteur de corrélation entre chaque valeur prédite et la valeur réelle correspondante dans les données de test montrent une prédiction fiable pour le revenu et la croissance de revenu Y2Y (table 10). La valeur absolue de l'erreur RMSE pour le revenu semble importante mais il faut la comparer à la valeur moyenne du revenu. Le taux R2 de détermination est très proche de 1, montrant une forte proximité entre la valeur prédite et la cible. Cela met en évidence la pertinence du modèle utilisé. Ce n'est pas le cas pour la prédiction de la surprise bien que les valeurs de mesure de reste d'erreurs sont faibles. Pour la surprise, comme la métrique R2 est négative, on peut en déduire que la régression linéaire est moins performante qu'un modèle qui attribuerait comme prédiction systématiquement la valeur moyenne de la surprise. Ces résultats correspondent aux anticipations présentées dans les diagrammes précédents où on observait une linéarité entre le revenu réel et le revenu estimé, la croissance Y2Y réelle et la croissance Y2Y estimée alors que ce n'est pas le cas pour la surprise.

De cette analyse, on peut observer que le revenu et la croissance se prédisent mieux que la surprise (calculée journalièrement). Les variables caractéristiques DayGap et Year EBIT Mean ne sont pour ainsi dire pas utilisés comme caractéristiques prédictives. En effet, DAYGAP ne donnent pas d'informations numériques pour la prédiction mais un indice du temps restant durant lequel les valeurs peuvent fluctuer. Comme indiqué précédemment, la variable Year EBIT Mean a été prise en compte mais il aurait été préférable de prendre la valeur EBIT trimestrielle. Dans la figure 13, on observe un coefficient négatif pour la croissance Y2Y lors de la prédiction de la surprise sur le jeu train 19 et test 20 qui peut s'interpréter comme une inversion de tendance entre la croissance et la surprise. Un coefficient nul indique que la prédiction ne dépend pas de cette variable. Théoriquement, la valeur absolue des coefficients indique l'importance des variables. Ainsi, on observe que la prédiction en revenu dépend principalement de la croissance et de la surprise Y2Y. Cependant, les coefficients de régression peuvent être trompeurs lorsque les données ne sont pas dans le même range de valeurs. C'est pourquoi une analyse de l'importance des variables caractéristiques a été réalisée dans la section suivante en effectuant un scaling et en utilisant la technique du gradient boosting.

4.4 Comparaison de l'importance des variables caractéristiques

Les différences d'échelle entre les variables caractéristiques utilisées dans la modélisation ne permettent pas d'identifier leurs importances relatives dans la prédiction de la cible. Il faut préalablement rapporter ces valeurs dans une échelle comparable en amont de la régression linéaire. La technique initialement utilisée MinMaxscaler met à l'échelle toutes les instances de données dans la plage [0, 1] ou [-1, 1] s'il existe des valeurs négatives dans l'ensemble de données. Un gros inconvénient de cette technique est que la présence de valeurs extrêmes (outliers) provoque une compression en un range très petit, des valeurs dans le

2ieme et 3ième quartile. La technique finalement utilisée RobustScaler permet de limiter l'impact des outliers. RobustScaler transforme la variable caractéristique en soustrayant la médiane, puis en la divisant par l'écart interquartile (valeur de 75 pourcent - valeur de 25 pourcent).

Apprentissage train_19_20, test_21

Cible	DayGap	Revenu - Mean	Season Growth Estimate	Year EBIT-Mean	Previous Season Surprise	Intercept	RMSE on test	R2 on test
Revenu	-6,29E+06	4,06E+09	1,22E+07	1,89E+07	2,41E+07	2,39E+09	11642438	0,99
Surprise	0,91	-0,80	0,39	2,63	14,97	-11,12	26.291	-0,02
Y2Y Revenu Growth	0,01	-0,01	3,03	0,03	0,21	-1,15	1.2298	0,96

FIGURE 14 – Tableau des coefficients du modèle de régression linéaire sur fichier d'apprentissage mis à l'échelle pour le jeu train 19-20 test 21

On observe sur la figure 14, que pour le revenu, la variable caractéristique avec le plus gros poids est la prédiction moyenne des experts, comme anticipé. Pour la surprise, c'est la surprise du trimestre identique de l'année précédente qui est privilégié et pour la croissance Y2Y du revenu, c'est l'estimé de croissance Y2Y. C'est ce qui était également attendu.

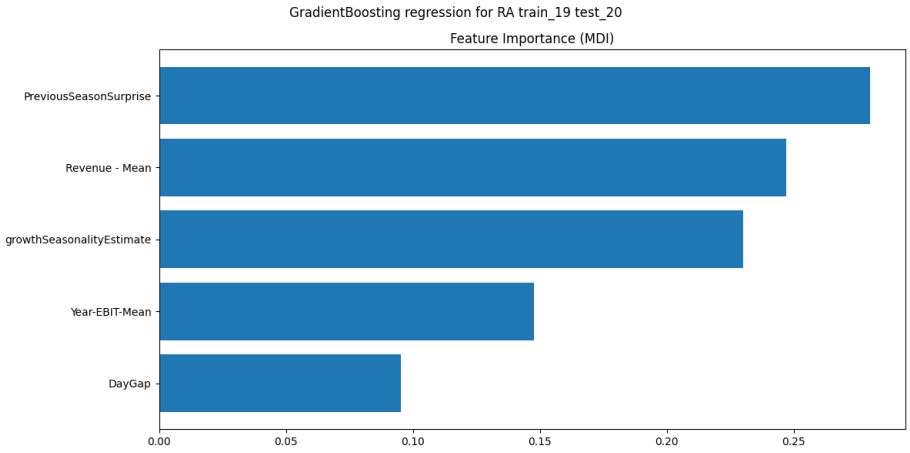

FIGURE 15 – Importance des variables caractéristiques avec une prédiction en revenu (RA=Revenu Actual) par GradientBoosting

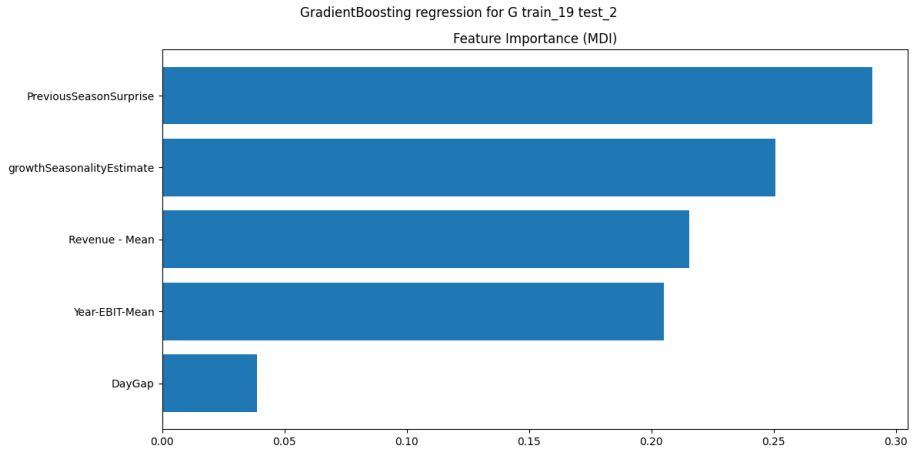

FIGURE 16 – Importance des variables caractéristiques avec une prédiction en croissance de revenu Y2Y par GradientBoosting

La méthode de gradient Boosting offre une autre technique de mesure de l'importance des variables caractéristiques (features) pour un exercice de prédiction. L'exercice (figure 15) a été réalisé sur le jeu d'apprentissage train 2019 test 2020 pour la prédiction du revenu et en croissance Y2Y. Il montre l'importance comme informations caractéristiques dans l'exercice de prédiction, de la surprise de l'année précédente, de l'estimé moyen en revenu des experts et de l'estimé en croissance Y2Y, dans une moindre mesure de l'estimé en EBIT annuel et du différentiel en jour entre la date de l'estimé par rapport à la fin de la période considérée. L'exercice (figure 16) correspond à la prédiction de la croissance de revenu. On peut observer que dans cette prédiction, l'importance de l'estimé en croissance de revenu prédomine l'estimé en revenu.

Cette technique de prédiction est consommatrice en temps avec nos volumes de données pour la prédiction de la surprise. Les tentatives pour cette prédiction ont tournés plus de 24 heures sans obtenir de résultats. De plus, seul le jeu d'apprentissage train 19, test 20 donne des résultats, les autres font saturer la mémoire de l'ordinateur.

En comparant avec l'approche de scaling, on observe pour ce seul résultat obtenu, que ici, la variable *Previous season surprise* prédomine tant pour la prédiction de revenu que pour celle de la croissance de revenu Y2Y. C'est étonnant par rapport à notre anticipation mais cela montre que la variable *Previous season surprise* a une force prédictive et peut être prise comme référence.

4.5 Comparaison des distributions des trois modèles de surprise

L'étape suivante est de construire les surprises à partir des prédictions en revenu (*Surprise by RA*) et en croissance de revenu Y2Y (*Surprise by G*). Ces trois surprises anticipées *Predicted Surprise*, *Surprise by Ra*, *Surprise by G* et la référence *previous season surprise* sont alors comparées à la surprise calculée sur le jeu de données de test *Computed Surprise*. La pertinence ou qualité de nos résultats sont mesurées par la métrique des anchored hit rates présentées ci-après. Dans cette section, nous présentons la comparaison de distribution de probabilité des variables. Bien que ces comparaisons ne nous permettent pas de conclure sur la qualité de la prédiction ou estimation de la surprise, elles donnent des indications qui peuvent nous aider à comprendre les différences systématiques.

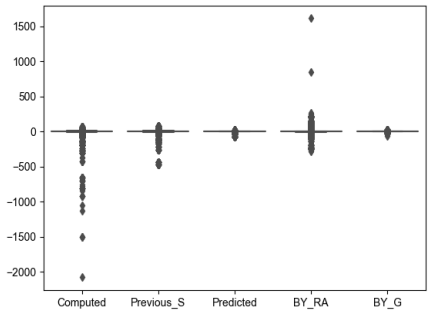

FIGURE 17 – Boxplots des surprises (Régression Linéaire, train 19-20 test 21)

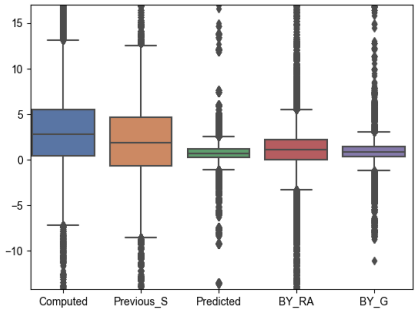

FIGURE 18 – Zoom sur Boxplots des surprises (Régression Linéaire, train 19-20 test 21)

La vue globale du boxplot, figure 17, montre que les valeurs extrêmes que présente Computed Surprise ne sont pas représentées dans la surprise prédite par apprentissage. La winsorisation utilisée avant apprentissage en est probablement la cause.

La comparaison en boxplot plus détaillé (figure 18) montre une distribution des trois surprises anticipées avec une plus petite variance autour de la moyenne, que la distribution de Computed surprise calculée journalièrement sur les données de test. La moyenne de la *surprise by RA* est proche de 1,2, celles de *surprise by G* et *Predicted Surprise* sont de 0,8 et 0,5 respectivement.

De ces résultats, aucune des distributions des surprises anticipées en régression linéaire ne ressemblent à celle de *Computed Surprise*. La surprise de la période Y2Y précédente *previous season Surprise* semble la plus proche.

La méthode en GradientBoosting donne une autre approche de prédiction que nous avons utilisé pour déterminer l'importance des variables caractéristiques. Vu les problèmes de charge, la prédiction n'a pu être réalisée que sur le revenu ainsi que la croissance et, de plus, uniquement sur le jeu de données train 19 et test 20. Comme présenté sur les figures 19 et 20, les résultats en prédiction donnent une distribution de la *surprise by RA* et la *surprise by G* fort éloignées de celle de *Computed Surprise*. La variance de la surprise anticipée par la prédiction de RA est très large alors que celle de la *surprise by G* est très étroite et fortement décalée en positif. En essayant de comprendre, on observe un domaine de valeurs pour la prédiction de croissance Y2Y fortement décorrélé des valeurs possibles, ce qui nous empêche de retenir ses résultats par la suite.

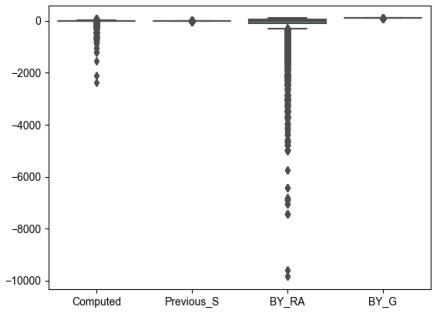

FIGURE 19 – Boxplots des surprises (Régression GradientBoosting, train 19 test 20)

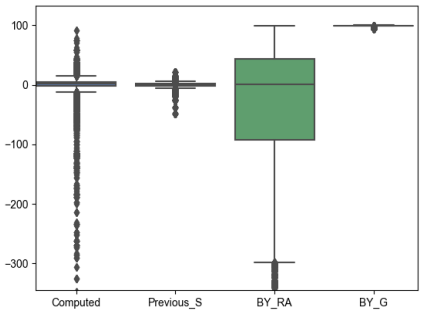

FIGURE 20 – Zoom sur Boxplots des surprises (Régression GradientBoosting, train 19 test 20)

4.6 Analyse de la force prédictive en Hit Rate

Cette section vise à comparer les signes des trois surprises anticipées et de la surprise référence (previous season surprise), à la surprise calculée. En préambule, il est important de prendre en compte que l'année 2020 est celle du début de la pandémie avec un gel complet de l'économie. L'année 2021 a dans une moindre mesure, également été marquée par la pandémie et a observé une certaine reprise en cours d'exercice.

Hit Rate	**Regr. lineaire, Train 19-20, Test 21**	**Regr. lineaire,Train 19, Test 20**	**Regr. lineaire,Train 20, Test 21**	**Gradient Boosting, Train 19, Test 20**
Référence				
Previous Season Surprise	62.78%	**55.09**%	62.78%	**55.09**%
Surprises Anticipées				
Predicted Surprise	71.47%	42.31%	**74.44**%	
Surprise by RA	66.42%	46.07%	69.59%	51,79%
Surprise by G	**72.18**%	45.75%	74.17%	

TABLE 11 – Pourcentage de Hit selon les méthodes de prédiction de la surprise en revenu

Les résultats repris dans la table 11 ne montrent pas une tendance cohérente

permettant de qualifier les modèles de prédiction. Le meilleur résultat est obtenu par un apprentissage de prédiction linéaire de la surprise sur l'année 2020 et un test sur l'année 2021. Cependant, chaque jeu d'apprentissage présente une méthode différente comme meilleure. La surprise établie par prédiction de croissance Y2Y est la meilleure pour le jeu train1920 test 20. On observe que dans le premier jeu et le troisième, les méthodes d'anticipations battent la référence et donc, que ces méthodes apportent de l'information supplémentaire à l'analyse décisionnelle d'achat/vente/conservation des actions des entreprises. Par contre, le jeu d'apprentissage en 2019 et test 2020 montre un recul de la qualité de prédiction de tous les estimateurs avec comme meilleur estimateur la référence *previous season surprise*. De plus, paradoxalement, en donnant plus de données d'apprentissage (années 2019 et 2020) pour un test sur la même année, on observe des résultats inférieurs quelque soit les méthodes d'anticipation. L'hypothèse pour expliquer ces deux constats est que la dynamique économique en 2019 ne correspond pas à celle de l'année 2020 utilisée en test, vu la crise due à la pandémie a faussé les modèles de prévision d'abord des analystes, ensuite de l'apprentissage.

4.7 Analyse de la force prédictive en Anchored Hit Rate

Anchored Hit Rate	**R. lineaire, Train 19-20, Test 21**	**R. lineaire, Train 19, Test 20**	**R. lineaire, Train 20, Test 21**	**Gradient Boosting, Train 19, Test 20**
Référence				
Previous Season Surprise	51.05%	50.99%	51.05%	50.99%
Prédiction				
Predicted Surprise	51.26%	51.00%	52.28%	
Surprise by RA	**53.05**%	50.18%	**53.26**%	**53.32**%
Surprise by G	50.47%	**51.2**%	50.79%	

TABLE 12 – Pourcentage de anchored Hit selon les méthodes de prédiction de la surprise en revenu

Après retrait de la médiane à chacune des variables permettant d'éviter le biais de surprises majoritairement positives, les résultats de la table 12 montrent en régression linéaire, un avantage pour le modèle de prédiction en revenu pour 2 jeux de test et un avantage très léger pour le modèle en

prédiction de la croissance pour le jeu train 19-test 20 par rapport au résultat de la référence. Comme pour les *hit rate*, le fait de disposer des données de deux années d'apprentissage, ne permet pas d'obtenir une meilleure prédiction pour l'année 21.

Pour le jeu de données (apprentissage 19 et test en 20), les résultats des estimateurs en regression linéaire sont tous en baisse, à l'exception, de manière surprenante, de l'estimateur par la prédiction de la croissance en revenu. Dans ce cas, l'estimateur par prédiction de revenu est le moins performant. Celui-ci dépend fortement, comme variable caractéristique, de l'estimation moyenne des analystes. L'estimateur de référence lui reste stable et ne dépasse que faiblement un estimateur théorique purement aléatoire qui attribuerait 50 pourcent de probabilité à une surprise positive et 50 pourcent à une surprise négative.

Il est difficile de se prononcer sur la force de prédiction de la technique en gradientBoosting puisqu'un seul jeu a pu être testé mais on observe pour cette méthode le meilleur anchored hit rate obtenu. Il faudrait néanmoins retravailler cette technique avec des jeux de données plus petits pour valider cette méthode.

En conclusion, on observe que dans les trois cas de régressions linéaires étudiés, la détermination de la surprise sur base de prédiction de valeurs économiques est préférable. De plus, selon la métrique anchored hit rate, la *Surprise by RA* est, généralement, la meilleure prédiction pour déterminer le signe de la surprise qui adviendra.

Etape 4 : Prédiction (régression linéaire (train_19 test_20)

---------clé---------

Date	cpy	Period End Date
21/05/20	A	31/07/20

Surprise predicted	Revenue Actual predicted	seasonal growth predicted
0,068023596	124404948	-0,015389888

Surprise _BY_RA	Surprise _BY_G
-0,583371884	0,2460442

anchored_ Surprise_ by_G	anchored_ Surprise_ by_RA	anchored_ Surprise_ predicted	anchored_ Previous Season Surprise	anchored_ Computed_ Surprise
0,377626577	-0,02906268	0,322903839	2,499217394	-0,853443037

hit_by_ Predicted_ Surprise	hit_by_ Previous_S_ Surprise	hit_by_RA	hit_by_G
1	1	0	1

Anchored_ hit_by_ Predicted_ Surprise	Anchored_ hit_by_ Previous_S_ Surprise	Anchored_ hit_by_RA	Anchored_ hit_by_G
0	0	1	0

FIGURE 21 – Résultats pour le cas exemple (train 19 test 20)

La figure 21 présente les résultats obtenus dans le cas exemple présenté lors de la création du journal. Le signe de la surprise obtenue après prédiction du revenu et ajustée par rapport à la médiane des valeurs correspond à celui de la variable *Computed Surprise* ajustée de la même façon.

5 Conclusion

Sur base des résultats obtenus, il est difficile de conclure qu'un modèle s'impose par rapport aux autres. Il est nécessaire d'augmenter les cas de tests pour conforter notre hypothèse que la détermination de la surprise après prédiction du revenu est préférable. En effet, pour deux jeux d'apprentissage et test, c'est ce même modèle qui s'impose en terme de 'anchored hit rate' excepté lors de la coupure due à la crise Covid (19-20) où la prédiction en croissance donne un faible avantage.

Dans nos trois cas temporels analysés, la prédiction directe de la surprise par apprentissage n'a pas donné de résultats probants.

Sous réserve de tests supplémentaires, on peut émettre comme hypothèse que dans les cycles économiques stables où les années d'apprentissage permettent au modèle de s'appuyer sur des estimatifs d'experts qui peuvent eux-même se fonder sur une certaine anticipation stable des revenus et des paramètres économiques, il pourrait être recommandé d'utiliser le modèle en prédiction de revenu pour déterminer la surprise. Par contre, lorsque des événements majeurs modifient globalement l'économie, ces estimations d'experts sont nettement moins fiables, ce qui perturbent fortement la pertinence de ce modèle au niveau de l'apprentissage.

Pour corroborer cette hypothèse, des analyses similaires devraient être réalisées par secteur d'activité. En effet, certains secteurs comme le pharmaceutique et l'e-commerce ont été moins impactés par la crise covid. Le fait de ne pas rencontrer cette même baisse de performance de nos estimateurs conforterait cette hypothèse.

Les prédictions ont été établies sur base de régression linéaire mais d'autres modèles plus complexes utilisant les réseaux neuronaux pourraient également être testés. De plus, il serait judicieux de réaliser ces mêmes tests non plus en utilisant trois dates fixes de séparation entre les données d'apprentissage et de tests mais par coupure glissante et de limiter l'espace temporel de test à quelques jours, ce qui permettrait d'obtenir beaucoup plus de jeu de tests tout en simulant le travail de gestionnaire de portefeuille. En effet, cela n'a pas beaucoup de sens de réaliser la prédiction à la dernière date du jeu de test, c'est à dire à un an des données d'apprentissage, autant utiliser comme apprentissage tout ce qui est connu à la date de la prédiction.

Il faudrait également modifier les erreurs que nous avons détectées, il aurait été préférable de travailler avec l'EBIT estimé trimestriel, de revoir le processus de création du journal pour reporter le nettoyage le plus tard et éviter les réductions de volumes en cascade. Cela permet d'améliorer le choix des variables caractéristiques et la quantité des données prises en compte, ce qui, avec le choix de modèle prédictif, correspond à des limites de notre approche.

Finalement, certaines préparations dans les données n'ont pas pu être exploitées, faute de temps. En effet, la préparation des données en un journal en séparant bien les données connues à la date du journal, des données cibles connues lors de la publication des résultats a nécessité un effort important pour éviter les *forward leak* ou la divulgation d'informations du futur. Dès lors, il serait intéressant de refaire ces mêmes analyses en séparant les tailles d'entreprises, les secteurs d'activité et la gouvernance sociétale des entreprises et d'interpréter si les différences des chiffres ont une explication rationnelle.

I Annexes

Les annexes décrivent les objectifs des différents programmes python, le séquencement utilisé pour la préparation du journal de données dans le programme python *Data preparation*, les différentes variables utilisées dans le journal, le résultat des prédictions sur les trois jeux de données. Elles contiennent aussi une analyse de pertinence des *anchored hit rate* sur base d'un regroupement par date ou par entreprise.

II Description des programmes en python

II.1 Data preparation.py

L'objectif est la réalisation d'un journal par quadrimestre d'entreprise simulant de manière chronologique l'obtention la valeur moyenne des estimés des analystes financiers tout en regroupant ces données avec les valeurs finales publiées pour ce quadrimestre.

Les données extraites de Refinitiv sont utilisées comme input par le programme. Les fichiers résultats sont le fichier journal output.xlsx dans le répertoire étapePréparation et les fichiers train.xlsx et test.xlsx dans le répertoire fichiersApprentissage.

Les paramètres de ce programme à adapter sont le chemin de répertoires et les variables de temps pour couper le journal et créer des fichiers d'apprentissage spécifiques.

II.2 DatacontentAnalysis.py

L'objectif est de donner les dimensions du contenu du journal et des fichiers d'apprentissages (nombre de lignes, nombre d'entreprises,nombre lignes par entreprise,nombre de date du journal, première date, dernière date, premier (test) / dernier (train) quadrimestre,nombre d'association entreprise - quadrimestre, nombre de jour par association entreprise - quadrimestre). Le chemin des répertoires utiles est à paramétrer dans le programme (fonction path())

II.3 SurpriseAnalysis.py

Comparaison de la surprise calculée à la date de publication par rapport à la surprise publiée par Refinitiv et Analyse du coefficient de corrélation de Pearson et Spearman. Les fichiers input sont `etapePreparation/journal_surprise.xlsx` et fichierDeRefinitiv/RevSurprise.xlsx.

II.4 cibleAnalysis.py

L'objectif est de présenter à partir du fichier journal (etapePreparation/output), les statistiques descriptives des variables cibles potentielles : Revenue Actual, Surprise (computedSurprise à la publication), ComputedSurprise (journalière), growthSeasonalityActual mais aussi d'analyser leur corrélation avec leurs variables estimateurs Revenue Mean, PreviousSeasonalSurprise, growthSeasonalityEstimate.

II.5 featuresAnalysis.py

L'objectif est de présenter à partir du fichier journal (etapePreparation/output), les statistiques descriptives des variables prédictives caractéristiques potentielles : DayGap, Revenue Mean, PreviousSeasonalSurprise, growthSeasonalityEstimate, Year-EBIT-Mean, previous season Revenue Actual, previous season EBIT.

II.6 featureimportance.py

Régression selon GradientBoosting sur un jeu d'apprentissage. Les trois cibles sont Revenue Actual, ComputedSurprise, GrowthSeasonalActual à partir des 5 variables prédictives : DayGap, Revenue - Mean, growthSeasonalityEstimate, Year-EBIT-Mean, PreviousSeasonSurprise. Présentation de l'importance des variables prédictives caractéristiques, de la mesure r2 , de la mesure MSE et des graphiques de distribution de probabilité des variables cibles et prédites du jeu de test. Sauvegarde de la prédiction dans le fichier `test_predicted_By_GBoosting`.

II.7 scaling.py

Regression linéaire sur un jeu d'apprentissage (train, test) dont les variables prédictives sont scalées selon soit le principe MinMax soit le principe robustScaler. Les trois cibles sont Revenue Actual, ComputedSurprise, GrowthSeasonalActual à partir des 5 variables prédictives : DayGap, Revenue - Mean, growthSeasonalityEstimate, Year-EBIT-Mean, PreviousSeasonSurprise. Présentation des coefficients de régression, de l'intercept, de la mesure r2, de la mesure MSE et des graphiques de distribution de probabilité des variables cibles et prédites du jeu de test.
Sauvegarde de la prédiction dans le fichier
"../fichierwithPredorHit/`test_predicted_by_train_scaled`.xlsx"

II.8 targets prediction.py

Regression linéaire multi-variée sur un jeu d'apprentissage pour les trois cibles Revenue Actual, ComputedSurprise, GrowthSeasonalActual à partir des 5 variables prédictives : DayGap, Revenue - Mean, growthSeasonalityEstimate, Year-EBIT-Mean, PreviousSeasonSurprise.

Présentation des coefficients de régression, de l'intercept, de la mesure r2, de la mesure MSE et des graphiques de distribution de probabilité des variables cibles et prédites du jeu de test. Sauvegarde de la prédiction dans le fichier `test_predicted.`

II.9 HitRateAnalysis.py

Construction des surprises prédites directement ou construites à partir de la prédiction de revenue ou de seasonal revenue growth. Construction des métriques hit et `anchored_hit` sur base des signes des surprises (prédites ou calculées à partir d'une prédiction versus la surprise calculée sur base de la publication. INPUT : `../fichierwithPredOrHit/test_predicted.xlsx` OUTPUTS : `../fichierwithPredOrHit/test_hit_rate.xlsx`

II.10 AnchoredHitRateAnalysis.py

Analyse des anchored Hit rates par des regroupements soit par date (histogramme du nombre de cpy avec anchored hit rate positif par date), soit par entreprise (histogramme du nombre de date avec anchored hit rate positif par entreprise).

III Description de la construction du journal de données

La construction du journal nécessite tout d'abord un nettoyage des données(pas de na, pas de nul) des différents fichiers obtenus de refinitiv.

III.1 Liens entre revenus et EBIT

FIGURE 22 – Établissement du lien entre revenu et EBIT

Tout d'abord, comme présenté à la figure 22, le fichier contenant le revenu publié (Actual) est rapproché de celui de l'EBIT publié, sur base de la clé constituée par l'entreprise et le quadrimestre concerné par les publications.

Dans une étape suivante, le traitement se base à partir du nouveau dataframe qui est pris en mémoire en triple pour pouvoir faire les liens avec le quadrimestre précédent et celui correspondant de l'année précédente.

III.2 Liens entre revenus ou EBIT par quadrimestre

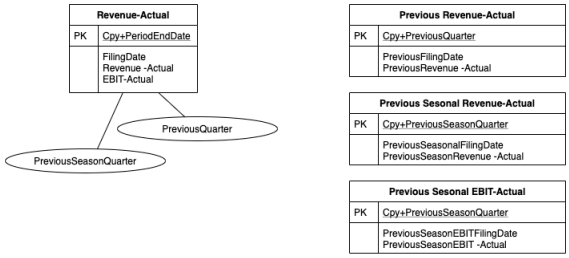

FIGURE 23 – Établissement des liens avec le revenu ou EBIT des périodes précédentes (directes et Y2Y)

Comme présenté à la figure 23, les noms de colonnes sont modifiés pour permettre les 2 merges sur base de la clé cpy + previousQuarter ou cpy + previousSeasonQuarter. Ces merges permettent de disposer du previousSeasonalRevenueActual et de la date PreviousFilingDate qui vont être utiles pour calculer le seasonal growth et pour établir le début des dates du journal pour chaque quadrimestre. De même, le fichier EBIT-Actual est copié en modifiant le nom des colonnes pour permettre d'obtenir l'EBIT du quadrimestre identique de l'an passé.

III.3 Création du journal

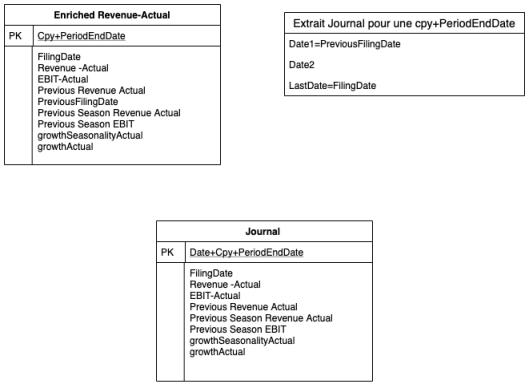

FIGURE 24 – Création du journal par date elligible

Comme présenté à la figure 24, pour chaque quadrimestre d'une entreprise du fichier Revenue Actual enrichi, on peut construire un journal partant de la date de publication du dernier quadrimestre et allant à la date de publication de ce quadrimestre. Les données de ce quadrimestre et cette entreprise sont alors recopiés sur chacune des dates correspondantes.

III.4 Lien avec les prévisions en revenu des analystes

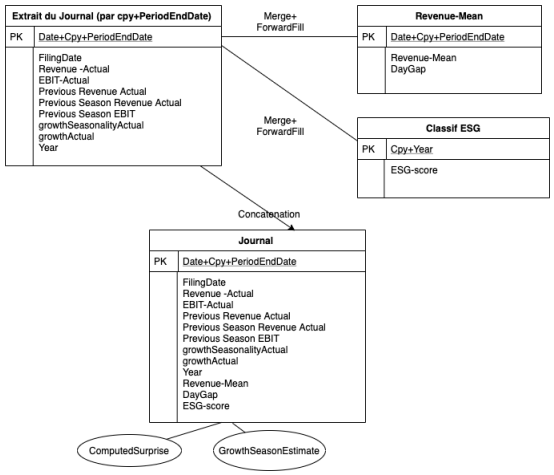

FIGURE 25 – Enrichissement du journal par les prévisions des analystes et le score ESG annuel

Comme présenté à la figure 25, les prévisions des analystes financiers sont associés à la date de. leur émission dans l'extrait du journal et une action de remplissage automatique vers l'avant est réalisée tant qu'une nouvelle prédiction des analystes n'est pas réalisée. Chaque extrait spécifique à une entreprise et un quadrimestre est alors ajouté au journal complet. Les colonnes "ComputedSurprise", "growthseasonalEstimate" peuvent ensuite être ajoutées.

III.5 Recherche de la surprise finale et création d'un journal des surprises

Sur base du journal ainsi obtenu, il est possible de le parcourir et ne retenir que la dernière ligne d'une section horizontale spécifique à un quadrimestre pour une entreprise particulière. Cette dernière ligne est alors ajoutée à un `journal_surprise` qui contient ainsi la surprise effective obtenue lors de la publication des résultat du quadrimestre. Un merge est effectué pour relier également la surprise du quadrimestre de l'année précédente à ce journal.

III.6 Lien avec la surprise finale et la surprise du quadrimestre correspondant l'année précédente

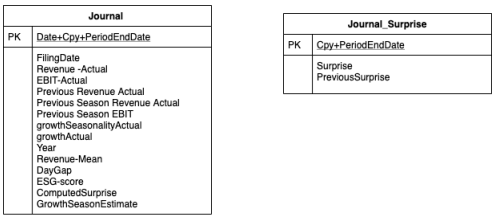

FIGURE 26 – Enrichissement du journal par la surprise et la surprise du quadrimestre Y2Y

L'enrichissement se poursuit en intégrant la surprise et la surprise de l'année précédente dans le dataframe journal comme illustré 26.

III.7 Lien avec la prévision EBIT des analystes

Un journal des prévisions EBIT des analystes est constitué entre deux dates spécifiées dans le programme. La technique de forwardfill et de concaténation des résultats obtenus par entreprise et quadrimestre est utilisée. Le journal enrichi à partir des revenus peut alors recevoir la prévision EBIT-Mean en utilisant la clé date+cpy+PeriodEndDate.

III.8 Classification des données quasi-statiques

Il reste à qualifier les entreprises en établissant trois attributs : secteur d'activité, taille (nombre d'employés) et ESG en opérant un regroupement simplifié

des valeurs des données quasi-statiques.

IV Description des variables du fichier journal et du fichier test

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
FilingDate	date de publication par l'entreprise du revenu et de l'EBIT
Revenu Actual	revenu publié par l'entreprise pour le quadrimestre considéré
growthActual	croissance du Revenu Actual par rapport à celui de la période précédente
growthSeasonalityActual	croissance du revenu par rapport à celui de la période Y2Y précédente
Previous revenueActual	revenu publié pour la période précédente
Revenue- Mean	Estimé moyen de revenu des experts connu à la date 'Date'
DayGap	Nombre de jours entre Period End Date et Date
growthSeasonalityEstimate	croissance du revenu estimé par rapport à celui de la période Y2Y précédente
growthEstimate	croissance du revenu estimé par rapport à celui de la période précédente
EBIT- Actual	EBIT publié par l'entreprise pour le quadrimestre considéré
Year-EBIT- Mean	Estimé moyen d'EBIT annuel établi par des experts et connu à la date 'Date'
Previous Season EBIT	EBIT publié par l'entreprise pour le quadrimestre Y2Y précédent

FIGURE 27 – Variables du journal

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
…	
ComputedSurprise	100*(Revenu estimé- Revenu Actual)/revenu Actual
CapComputedSurprise	ComputedSurprise windsorisée à 0.05%, 0.05%
Surprise	Valeur de ComputedSurprise à la date de publication de l'entreprise
Previous seasonal surprise	Surprise de la période Y2Y précédente

FIGURE 28 – Variables du journal (part2)

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
…	
PreviousSeasonQuarter	Quadrimestre Y2Y précédent identifié par sa date de Period End Date
PreviousQuarter	Quadrimestre précédent identifié par sa date de Period End Date
Year	Année en cours à la date 'Date'
ESG Score	Valeur ESG établie pour l'année en cours
Business Sector	Domaine d'activité de l'entreprise (avril 2022)
Number of employees	Nombre d'employés (avril 2022)
Size	Small(<1K #employés), medium(<10K), big(<100K), , huge (avril 2022)
ESG	toImprove (<25),neutral (<50),good(<>75),veryGood
Business	Regroupement de domaine d'activité (avril 2022)

FIGURE 29 – Variables du journal (part3)

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
…	
predictedSurprise	Surprise établie par prédiction après apprentissage
Seasonal growth predicted	Croissance Y2Y de revenu prédite
Revenue Actual predicted	Revenu prédit
RA_DEDUCED_BY_G_Pred	=(test['previous season Revenue Actual']+(test['seasonal growth predicted']*test['previous season Revenue Actual']))
Surprise_DEDUCED_BY_RA_Pred	Surprise déduite par la prédiction de Revenu =(100*(test['Revenue Actual predicted']- test['Revenu - Mean']) /test['Revenue Actual predicted']))
Surprise_DEDUCED_BY_G_Pred	Surprise déduite par la prédiction de croissance de revenu =(100*(test['RA_DEDUCED_BY_G_Pred']-test['Revenue - Mean'])/test['RA_DEDUCED_BY_G_Pred']))

FIGURE 30 – Variables ajoutées au fichier test

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
…	
Hit_by_Previous_S_S	True if ComputedSurprise et Previous seasonal surprise sont de même signe
Hit_by_pred	True if ComputedSurprise et Surprise predicted sont de même signe
Hit_by_RA_pred	True if ComputedSurprise et Surprise_deduced_by_RA_pred sont de même signe
Hit_by_G_pred	True if ComputedSurprise et Surprise_deduced_by_G_pred sont de même signe

FIGURE 31 – Variables ajoutées au fichier test (part2)

Variables	Description
Date	date courante pour la journalisation
[Cpy-Period End Date]	clé d'identification de l'entreprise et de la fin de période du quadrimestre suivi
…	
anchored_Surprise	ComputedSurprise - Mean(ComputedSurprise)
anchored_PreviousSeasonSurprise	PreviousSeasonSurprise - Mean(PreviousSeasonSurprise)
anchored_Surprise_predicted	Surprise predicted - Mean(Surprise predicted)
anchored_Surprise_by_RA	Surprise_DEDUCED_BY_RA_Pred - sa moyenne
anchored_Surprise_by_G	Surprise_DEDUCED_BY_G_Pred - sa moyenne
Anchored_hit_by_Previous	True if anchored_Surprise et anchored_PreviousSeasonSurprise sont de même signe
Anchored_hit_rate	True if anchored_Surprise et anchored_Surprise_predicted sont de même signe
Anchored_hit_by_RA_pred	True if anchored_Surprise et anchored_Surprise_by_RA sont de même signe
Anchored_hit_by_G_pred	True if anchored_Surprise et anchored_Surprise_by_G sont de même signe

FIGURE 32 – Variables ajoutées au fichier test (part 3)

V Établissement des surprises pour les différents jeux d'apprentissage

Apprentissage: train_19_20 , test_21							
	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
Prevision							
Surprise_DEDUCED_BY_RA_Pred	1,21	16,48	-280,64	0,01	1,15	2,20	1615,46
Surprise_DEDUCED_BY_G_Pred	0,86	2,02	-67,28	0,38	0,89	1,44	28,65
Predicted Surprise	0,45	3,67	-73,58	0,25	0,68	1,16	20,56
PreviousSeasonSurprise	-0,01	23,48	-475,26	-0,63	1,85	4,66	75,23
Cible							
ComputedSurprise (daily)	1,95	25,97	-2080,11	0,45	2,77	5,51	70,48
Surprise (end of Period)	1,97	20,29	-659,00	0,45	2,50	5,02	70,39

FIGURE 33 – Metriques des valeurs de surprises prédites

Apprentissage :train_19 , test_20							
	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
Surprise_DEDUCED_BY_RA_Pred	-2,68	63,89	-7791,16	-2,20	-0,55	0,10	528,53
Surprise_DEDUCED_BY_G_Pred	-0,11	1,48	-28,97	-0,35	-0,13	0,09	84,25
Surprise predicted	-0,27	0,53	-6,60	-0,44	-0,25	-0,04	2,48
PreviousSeasonSurprise	0,06	3,84	-48,36	-1,04	0,25	1,66	21,50
ComputedSurprise (daily)	-1,24	34,16	-2359,88	-1,89	1,62	5,11	91,31
Surprise (end of Period)	0,08	22,06	-475,26	-0,67	1,85	4,58	75,23

FIGURE 34 – Metriques des valeurs de surprises prédites

Apprentissage :train_20 , test_21							
	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
Surprise_DEDUCED_BY_RA_Pred	1,70	81,93	-10427,90	0,85	2,26	3,69	10027,71
Surprise_DEDUCED_BY_G_Pred	1,96	6,76	-235,25	1,21	2,13	3,06	415,91
Surprise predicted	1,86	10,04	-106,66	1,00	1,81	2,68	354,50
PreviousSeasonSurprise	-0,01	23,48	-475,26	-0,63	1,85	4,66	75,23
ComputedSurprise (daily)	1,95	25,97	-2080,11	0,45	2,77	5,51	70,48
Surprise (end of Period)	1,97	20,29	-659,00	0,45	2,50	5,02	70,39

FIGURE 35 – Metriques des valeurs de surprises prédites

Références

[1] John Alberg and Zachary C Lipton. Improving factor-based quantitative investing by forecasting company fundamentals. *arXiv preprint arXiv :1711.04837*, 2017.

[2] Jose I Alvarado, Lindsay C Clark, and Jose A Gutierrez. Stock performance subsequent to combinations in quarterly revenue surprise, earnings surprise, guidance, valuation, and report time. *Journal of Economics and Finance*, 45(1) :95–117, 2021.

[3] Charles Becque. forecasting earning surprises with machine learning. https ://towardsdatascience.com/forecasting-earning-surprises-with-machine-learning-68b2f2318936, 2019.

[4] Stephen Brown, Stephen A. Hillegeist, and Kin Lo. The effect of earnings surprises on information asymmetry. *Journal of Accounting & Economics*, 47, 2009.

[5] Marcos Lopez de Prado. *Advances in financial machine learning.* John Wiley & Sons, 2018.

[6] Yonca Ertimur, Joshua Livnat, and Minna Martikainen. Differential market reactions to revenue and expense surprises. *Review of Accounting Studies*, 8 :185–211, 2003.

[7] Narasimhan Jegadeesh and Joshua Livnat. Post-earnings-announcement drift : The role of revenue surprises. *Financial Analysts Journal*, 62, 05 2006. doi:10.2469/faj.v62.n2.4081.

[8] Joseph D. Piotroski. "value investing : The use of historical financial statement information to separate winners from losers.". *Journal of Accounting Research, vol. 38, 2000, pp. 1–41. JSTOR,.* doi:https://doi.org/10.2307/2672906.

[9] James Wahlen and Matthew Wieland. Can financial statement analysis beat consensus analysts' recommendations ? *Review of Accounting Studies*, 16 :89–115, 03 2010. doi:10.1007/s11142-010-9124-5.

[10] Lingling Zheng Xuemin (Sterling) Yan. Fundamental analysis and the cross-section of stock returns : A data-mining approach. *The Review of Financial Studies*, Volume 30, Issue 4 :Pages 1382–1423, April 2017. doi:https://doi.org/10.1093/rfs/hhx001.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm