Collot Laurène Sylla Jacques Axel M2 MBFA GDA

## Prédire le cours d'une action d'une entreprise à l'aide de méthodes de Machine Learning

L'investissement en bourse peut sembler intimidant pour de nombreux investisseurs, surtout lorsque nous prenons en compte les nombreux facteurs qui influencent les prix des actions sur les marchés financiers. Les économies mondiales étant en constante évolution, il est difficile de prévoir l'avenir des actions en fonction des politiques gouvernementales, des tendances du marché et des fluctuations de la monnaie. Les acteurs financiers utilisent ainsi couramment des variables quantitatives pour prédire les cours des actions, en partie grâce à des méthodes de Machine Learning. Cependant se concentrer uniquement sur les données quantitatives peut faire courir le risque d'ignorer les risques extra-financiers liés à un investissement.

L'importance de ses variables extra-financières dans les stratégies d'investissement boursier a été mise en lumière récemment à travers des événements tels que la pandémie de COVID-19 et le scandale Orpea. Il semble donc plus que jamais essentiel de considérer ses critères tels que la réputation de l'entreprise, la perception des consommateurs et les pratiques ESG lors de la prévision des prix des actions car bien que trop souvent ignorés, ces facteurs sont cruciaux pour prédire avec précision les prix des actions d'une entreprise et limiter les risques de chute de ces derniers.

C'est pourquoi, dans un monde en constante évolution, il est important de se poser la question de l'opportunité d'intégrer ces critères extra-financiers dans les modèles de Machine Learning. Ainsi est-ce que la prise en compte de facteurs tels que la réputation d'une entreprise, la perception des consommateurs ou les pratiques ESG est pertinente pour prédire correctement les prix des actions, ou ces variables, souvent difficiles à quantifier, ne sont-elles que des opinions et subjectivités pour les investisseurs ?

Dans notre étude, nous examinerons l'intégration de facteurs financiers, économiques et extrafinanciers dans la prédiction des prix des actions sur trois entreprises américaines bien établies dans leur secteur respectif : Microsoft dans la technologie, Exxon Mobil dans l'énergie, et JP Morgan dans la finance. Nous tenterons de développer un modèle qui peut prévoir les fluctuations du cours de bourse en utilisant une combinaison de variables quantitatives, telles que l'inflation américaine et la dette à long terme, qui sont largement utilisées dans l'industrie financière, ainsi qu'une variable extra-financière, à savoir le sentiment des consommateurs envers chaque entreprise. En incluant ces variables distinctes, nous visons à améliorer la précision de nos prévisions de prix des actions.

#### Sommaire

- 1. Récupération des données financières et création d'un data set
- 2. LSTM (long short-term memory)
- 3. Random Forest
- 4. Conclusion
- 5. Annexes

#### 1. Récupération des données financières et création d'un data set

Nous étudions trois secteurs différents : le secteur de l'énergie, de la technologie et de la finance, pour voir comment prédire une action d'une entreprise et surtout voir les différents résultats selon les secteurs.

Nous allons prendre le cours ajusté de clôture de l'action pour trois entreprises différentes : ExxonMobile, Microsoft, JP Morgan. Pour ces trois entreprises nous allons utiliser des données journalières allant du 31/12/2012 au début de l'année 2023 (car actualisation sur Alpha Vantage au cours d'aujourd'hui) : la période d'étude est d'un peu plus de 10 ans. Pour chaque cours d'actions, nous avons donc environ 2540 lignes et 1 colonne.

Ensuite, nous avons une variable qui est la dette long terme de chaque entreprise étudiée, prise trimestriellement et sur une durée allant du 31/12/2017 au 01/09/2022, sauf Microsoft qui a déjà publié les chiffres de la dette long terme pour le premier trimestre de 2023 (et qui du coup a des données qui commencent un trimestre de retard par rapport aux autres : le 31/03/2018).

La troisième variable qui va être utilisée durant cette étude est l'inflation (CPI) américaine qui est mensuelle et qui recouvre une période de janvier 2013 à décembre 2022 : ce qui nous fait 120 données.

Pour finir, la dernière variable utilisée est le sentiment moyen du consommateur qui est mensuelle et qui est créé à partir de tendance de google trend sur certains mots. Il s'agit de la tendance aux Etats-Unis du nom de l'entreprise concernée, et de 2 mots clés évoquant son secteur. Nous téléchargeons le résultat obtenu en fichier .csv que nous lisons dans python. Nous répétons ceci pour chacune des entreprises.

Nous avons récupéré les données sur Alpha Vantage à l'aide d'un API et nous avons ensuite concaténé les différentes variables dans un dataset sur python. Le dataset constitué contient de nombreux NaN en raison de la différence de fréquence des variables, nous devons les combler. Nous avons le choix entre remplacer les Nan par 0 ce qui biaiserai l'application du machine learning ou bien de remplir les NaN par la même donnée jusqu'à la suivante, ce que nous faisons. Nous aurons donc le même chiffre de l'inflation pour tout le mois. Nous obtenons alors le datasetclean.

Nous n'avons pas la même fréquence sur les variables prises et on peut remarquer qu'il y a des Nan. Les NaN dans la valeur des actions signifient qu'il y a un week-end car nous avons des données seulement quand les marchés financiers sont ouverts donc en semaine. Le 30 juin 2018 est un bon exemple car on a eu la valeur de la dette ce jour-là, le rapport trimestriel des entreprises est publié, donc nous avons la valeur de la dette qui apparait un samedi mais il n'y a en revanche aucune données pour les actions car le samedi les marchés financiers sont fermés.

Autre exemple, le dimanche 1<sup>er</sup> juillet 2018 était un dimanche, les marchés financiers étaient encore fermés ce qui veut dire que nous n'avons pas de données pour les actions. Cependant c'est le premier du mois donc l'inflation du mois de juillet vient se comptabiliser sur ce jourlà, (le 2 juillet 2018 qui était un lundi, nous avons à nouveau les cours de l'action mais pas de l'inflation car elle a été donnée le 1er du mois, elle est mensuelle).

# 2. <u>LSTM (long short-term memory) pour prédire les actions de nos</u> entreprises

Le LSTM (Long Short-Term Memory) est un type de réseau de neurones artificiels développé pour traiter les séquences de données, comme les séries chronologiques, les textes ou les signaux audio, la génération de texte et la classification de séquences. Il est également souvent utilisé en conjonction avec d'autres réseaux de neurones pour résoudre des problèmes plus complexes.

Cette technique a été conçue pour résoudre les limitations des réseaux de neurones traditionnels en matière de traitement de la mémoire à long terme, qui est l'un de leur problème majeur. Les réseaux de neurones traditionnels ont en effet tendance à oublier les informations anciennes à mesure qu'ils apprennent de nouvelles informations. Le LSTM remédie à ce problème en introduisant une mémoire interne appelée "cellule" qui peut stocker et contrôler l'accès à des informations à long terme.

Les cellules LSTM sont alors reliées à des entrées, des sorties ainsi qu'à une boucle interne. Grâce à cette boucle interne, la cellule peut conserver les informations importantes qui lui ont été transmises durant de plus longues périodes. Les entrées et les sorties permettent alors de contrôler ce qui est stocké dans la mémoire. Les entrées peuvent ainsi décider de garder ou non des informations, tandis que les sorties peuvent déterminer ce qui est transmis au réseau suivant. Les avantages de l'utilisation d'un réseau LSTM sont nombreux et peuvent varier en fonction de l'application spécifique. En général, le LSTM peut améliorer la performance de la prédiction notamment grâce sa capacité à mémoriser les données à long terme. Cela est particulièrement utile dans le cadre des séries chronologiques qui ont une dépendance temporelle, telles que les données financières.

Un autre avantage important est qu'ils peuvent mieux gérer les problèmes de surapprentissage en raison de leur capacité à contrôler l'effacement et l'écriture de données dans leurs cellules de mémoire.

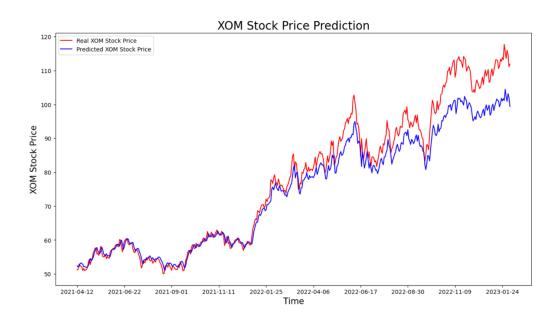
Cependant, il y a également des inconvénients à utiliser un réseau LSTM. L'un d'entre eux est que leur performance peut être affectée par le choix des hyperparamètres, tels que le nombre de couches cachées et la taille de la mémoire, ce qui peut rendre difficile la sélection des paramètres optimaux pour une tâche donnée. L'entraînement des réseaux LSTM peut également s'avérer lent en raison de leur grande taille et de leur capacité à mémoriser de grandes quantités de données. Enfin il peut exister des problèmes d'overfiting.

Concernant notre étude, nous importons au préalable les cours ajustés de clôture journalière à partir d'Alpha Vantage pour ExxonMobil, Microsoft et JP Morgan. On inverse l'ordre de la série temporelle puis on la divise en données d'apprentissage et en données de test. Comme il y a 2540 lignes dans la série temporelle et que nous voulons prédire 459 valeurs (soit 1,5 années de valeurs représentées sur un graphique), nous faisons la division entre training et test à la ligne 2081. Ceci représente une taille de test de 18% (=459/2540).

Il est nécessaire de normaliser les données à l'aide du MinMaxScaler avant de calibrer le modèle cela boostera la performance. Nous construisons l'input X train puis construisons le modèle LSTM avec 50 neurones et 4 hidden layers. Au final, nous assignerons un neurone dans l'ouput layers pour prédire le cours de l'action normalisé. Nous utiliserons le MSE loss function et l'Adam stochastic gradient descent optimizer.

Ensuite nous préparons les données de tests, puis nous faisons des prédictions en utilisant ces données de test. Nous visualisons les résultats sur un graphique qui compare le cours réel de l'action et la prédiction faite par le LSTM au cours de l'année demi passée.

Évolution du cours réel de l'action par rapport à la prédiction du cours à l'aide du LSTM pour Exxon Mobil

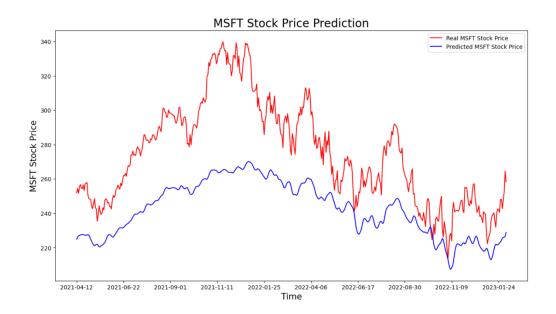


Pour ExxonMobil, le Mean Absolute percentage error est de 4,01% ce qui est très bas et donc satisfaisant, le modèle a très bien prédit la plupart des variations du cours de l'action.

L'écart entre les deux courbes est plus visible depuis l'année 2022. Nous pouvons voir que le modèle a prédit des valeurs plus basses comparées aux valeurs prédites de l'action Exxon Mobil qui a augmenté depuis la crise énergétique.

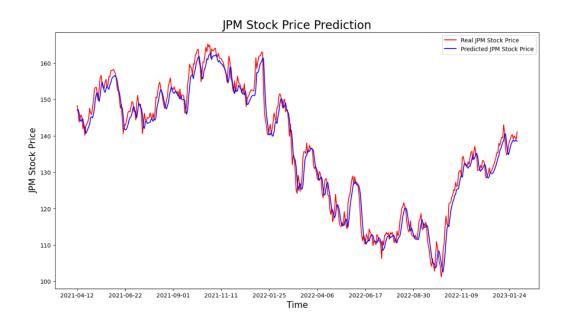
Évolution du cours réel de l'action par rapport à la prédiction du cours à l'aide du LSTM

pour Microsoft



Pour Microsoft, le Mean Absolute percentage error est de 11,27% ce qui est plus élevé que pour ExxonMobil et donc moins satisfaisant, le modèle a moins bien prédit les variations du cours de l'action Microsoft au cours de l'année demi passée. En effet, le modèle a prédit des valeurs beaucoup plus basses que les valeurs réelles, qui se sont envolées au dernier trimestre 2021.

## Évolution du cours réel de l'action par rapport à la prédiction du cours à l'aide du LSTM pour JP Morgan



Pour JP Morgan, le Mean Absolute percentage error est de 1,59% ce qui est négligeable et extrêmement satisfaisant, le modèle a vraiment bien prédit les variations du cours de l'action. Les deux courbes se superposent presque. Nous voyons que l'évolution du cours de l'action est assez cyclique et que le LSTM est performant pour prédire ce type de schéma, cette hypothèse semble aussi valide pour ExxonMobil qui fait partie du secteur de l'énergie étant une industrie cyclique.

#### 3. Random Forest

Le Random Forest est un algorithme d'apprentissage automatique très populaire, utilisé pour les tâches de classification et de régression. Il utilise une combinaison d'arbres de décision pour prédire une variable cible et se base sur la technologie développée dans les méthodes des arbres de décision.

Sur l'aspect technique les Random Forest utilise la force de plusieurs Décision Tree pour améliorer la précision de leur prédiction. Il se base ainsi sur des tests associés à des nœuds dans chaque arbre de décision. À chaque nœud, l'algorithme calcule le gain d'information pour chaque feature et choisit la feature qui maximise ce gain pour déterminer la question à poser. Les réponses à chaque question (oui ou non) orientant donc l'algorithme vers une branche spécifique.

Chaque arbre dans l'ensemble est donc formé en utilisant un sous-ensemble aléatoire des données d'entraînement, appelé "bootstrapped data", qui est créé en sélectionnant aléatoirement des échantillons de la base de données originale. Les arbres sont ensuite entraînés sur un sous-ensemble aléatoire du dataset et donne un résultat qui est ensuite combiné avec les résultats des autres arbres pour donner une réponse finale majoritaire. C'est ce qu'on appelle la méthode de bagging. La prédiction finale est effectuée en utilisant un vote majoritaire des prédictions de chaque arbre.

Les points forts du Random Forest sont essentiellement sa capacité à gérer les données manquantes, à prévenir le sur-apprentissage mais également à produire une bonne estimation de la variable cible. De plus, il peut fournir une estimation de l'importance des variables, ce qui peut être utile pour la sélection des variables et l'analyse de la caractéristique. Cependant, Random Forest possède également certaines limites. En effet l'algorithme peut être lent lorsqu'il traite de grandes quantités de données et peut ne pas être adapté pour les problèmes de classification binaire. Les résultats peuvent être difficiles à interpréter et le nombre d'arbres peut affecter les performances de l'algorithme.

Au préalable pour réaliser notre Random Forest qui va nous aider à prédire le cours de l'action de chaque entreprise, nous allons convertir notre datasetclean en array et prendre seulement les variables associées à cette entreprise.

Pour l'entreprise Exxon Mobil, nous avons 3 variables explicatives (X) pouvant expliquer le cours ajusté l'action de celle-ci : l'inflation (CPI) mensuelle, la dette long terme de Exxon Mobil (trimestrielle) et la moyenne du sentiment du consommateur (mensuelle). Nous faisons de même pour Microsoft et JP Morgan en prenant l'inflation, leur dette long terme et le sentiment consommateur moyen.

Le cours de l'action de l'entreprise concernée est la variable à prédire, y est la variable à expliquer qui se divise en y\_train et y\_test. En effet, on divise chaque variable X et y en sous échantillon, avec 70% des données qui sont les données d'apprentissage (X\_train ou y\_train) et 30% qui correspondent aux données de test (X test ou y test).

Nous entrainons le modèle sur l'échantillon de training et on performe les prédictions sur l'échantillon de test. Nous prenons 80 estimations ce qui signifie que notre simulation Random Forest contient 80 arbres de décision aléatoire. Ensuite, on calibre le modèle à partir des X\_train et y\_train. A partir du X\_test nous prédisons le cours ajusté de l'action Exxon mobil (y\_pred). y\_test et y\_pred ont une dimension de (777,1), ce qui est cohérent.

Étant donné que la taille du test correspond à 30% de la taille du dataset originel alors 30% de 2590 (nombre de lignes du dataset) nous donne 777.

Nous étudioins avec quelle fréquence le « regressor Random Forest » est correct. Concernant le score du modèle (rf\_model.score) de Exxon Mobil, la pertinence d'entraînement (training accuracy) est de 97,53% et celle du modèle 97,86%. Pour Microsoft, nous avons un training accuracy de 99,76% et un model accuracy de 99,72%. JP Morgan a un training accuracy de 99,36% et un model accuracy est de 99,08%.

Le training accuracy mesure la précision du modèle lorsqu'il est entraîné sur des données d'entraînement. La pertinence du modèle (model accuracy) mesure la précision du modèle sur les données qui n'ont pas été utilisées pour l'entraînement, donc sur les données de test.

Le training accuracy peut être très élevé, mais cela ne signifie pas nécessairement que le modèle aura une bonne performance. Un modèle sur-entraîné (overfitting) peut avoir un très bon training accuracy mais une mauvaise model accuracy. Il est important de mesurer les deux pour évaluer la performance réelle du modèle.

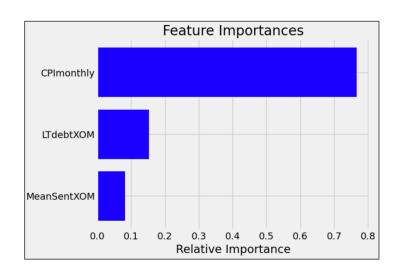
Pour les trois entreprises le training accuracy est élevé, ce qui veut dire que notre modèle est capable d'apprendre des données d'entraînement. Le model accuracy est aussi élevée et très proche de celle d'entrainement ce qui signifie que le modèle généralise bien aux nouvelles données. L'écart entre l'accuracy d'entrainement et l'accuracy du modèle est très faible ce qui prouve que nous n'avons pas d'overfitting ou d'underfitting. Ici nous avons donc un modèle fiable.

La Mean Absolute Error (MAE) est une métrique couramment utilisée pour évaluer les erreurs dans les modèles de prévision. Elle mesure la différence absolue moyenne entre les valeurs prévues et les valeurs réelles pour une série de prévisions. Elle donne une mesure simple et intuitive de la précision des prévisions. Dans notre étude, nous avons une relativement petite MAE pour les trois entreprises qui est de 2,15% pour Exxon Mobil et Microsoft, et de 2,51% pour JP Morgan. Cela veut dire que le modèle à de bonnes performances car plus la MAE est petite, plus les performances sont meilleures. Notre modèle a des prévisions proches des valeurs réelles. Cependant, il est important de noter que la MAE ne prend pas en compte la direction des erreurs (sous ou sur-estimation), seulement leur magnitude absolue.

Maintenant, on veut savoir quelle est l'importance de chacune des variables explicatives (CPI, dette, sentiment moyen consommateur) dans la prédiction du cours ajusté de clôture de l'action de nos différentes entreprises.

Pour Exxon Mobil, l'inflation (CPI) mensuelle est la variable la plus importante elle représente presque 76,58% dans l'importance relative du prix de l'action de Exxon Mobil. La dette de long terme représente 15,28% de l'importance relative comparé avec les autres variables. La moyenne du sentiment du consommateur a une importance relative de 8,14%.

Graphique de l'importance relative des variables explicatives dans la prévision du cours de l'action Exxon Mobil



Il parait logique que l'inflation ait une grande importance dans la prédiction du cours de l'action de Exxon Mobil car c'est une entreprise du secteur de l'énergie qui produit du pétrole. Par exemple, lorsqu'en 2022 le conflit russo-ukrainien a démarré et il y a eu un choc d'offre dû à l'augmentation du prix du pétrole, cela a contribué à augmenter l'inflation et par la même occasion le cours de l'action de Exxon Mobil. Donc quand l'inflation augmente, via la hausse du baril de pétrole, la valeur d'Exxon Mobil augmente sur les marchés financiers.

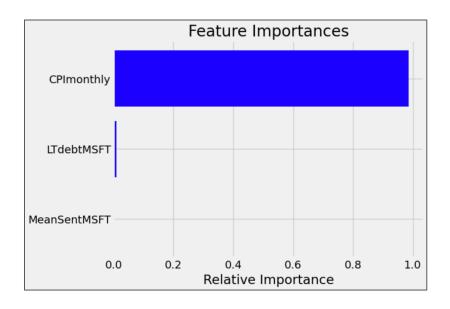
En 2022 avec la crise énergétique, il est logique de penser que le sentiment d'Exxon Mobil sur google trend a augmenté.

Pour essayer le modèle, on fait une prévision du cours de l'action d'Exxon Mobil pour le jour d'après : il suffit d'entrer des inputs pour les 3 variables explicatives et le modèle nous prédit le cours de clôture ajusté d'Exxon Mobil pour le jour suivant.

On suppose que nos trois variables explicatives restent les mêmes pour le jour d'après car elles sont mensuelles et trimestrielles. La valeur obtenue (prédite par notre modèle Random Forest) est de 112,07 de l'action ce qui est cohérent avec la dernière valeur du cours de l'action qui était de 111,89. Nous pouvons supposer que le mois suivant l'inflation diminuera légèrement ainsi que le sentiment moyen du consommateur pour Exxon mobil. En revanche la dette qui est trimestrielle reste la même jusqu'au 31 mars 2023. Comme attendu la valeur du cours de l'action prédite est en baisse pour le mois d'après, notre modèle Random Forest prédit 93,82 comme cours de l'action.

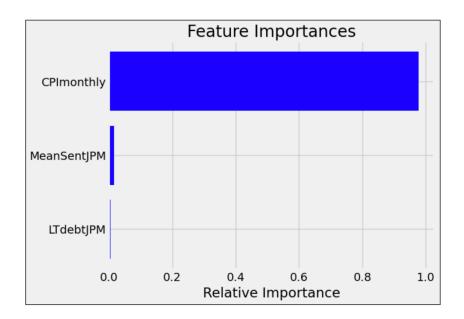
Au sujet de l'importance relative des variables dans la prévision du cours de l'action Microsoft nous avons l'inflation mensuelle qui représente 98,59% de l'importance relative de l'action Microsoft. La dette de long terme est de 0,98% et la moyenne des sentiments du consommateur est de 0,43%, dans l'importance relative. Nous avons clairement ici une inflation qui monopolise l'influence, c'est elle qui impacte le plus sur la prévision du cours de l'action Microsoft.

Graphique de l'importance relative des variables explicatives dans la prévision du cours de l'action Microsoft



Concernant JP Morgan les résultats du poids selon les variables de l'importance relative à la prévision du cours de l'action est très similaire aux chiffres de Microsoft avec une inflation qui domine (97,8%).

Graphique de l'importance relative des variables explicatives dans la prévision du cours de l'action JP Morgan



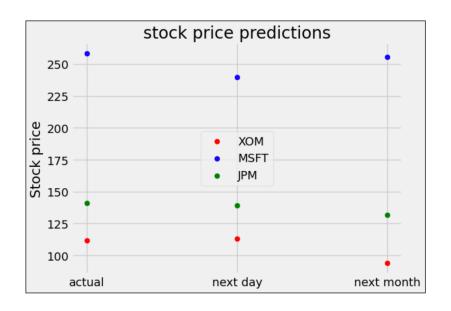
JP Morgan et Microsoft dépendent tous les deux beaucoup de l'inflation, quand on fait une prédiction au jour suivant ou au mois suivant, de monter ou diminuer l'inflation c'est ce mouvement de l'inflation qui va être déterminant du cours de l'action.

Concernant le domaine bancaire quand l'inflation augmente, les taux d'intérêts directeurs vont augmenter et donc les taux d'intérêts des prêts, ce qui va plus rémunérer les prêts grâce à l'augmentation du taux d'intérêt. Ce mécanisme va par conséquent augmenter la rémunération de la banque et donc son cours de l'action. La valeur obtenue du cours de l'action de JP Morgan est de 139,17 ce qui est cohérent avec la dernière valeur du cours de l'action mais nous avons une légère augmentation au mois suivant avec comme cours 131,58 ; car nous anticipons une baisse de l'inflation ce qui ferait baisser le cours de JP Morgan.

Pour Microsoft le mécanisme est à sens inverse, c'est justement quand l'inflation va augmenter que le cours de l'action Microsoft va diminuer car par exemple cela coûte plus cher de manufacturer un ordinateur quand l'inflation est élevée. Donc l'augmentation de ces coûts ne va pas être bénéfique pour l'entreprise Microsoft qui va devoir payer plus cher la production de ces produits et qui par conséquence va venir baisser le cours de son action. La valeur obtenue du cours de l'action de Microsoft est de 239,77 ce qui est cohérent avec la dernière valeur du cours de l'action mais nous avons une légère augmentation au mois suivant avec comme cours 255,77 ; car nous anticipons une baisse de l'inflation ce qui ferait augmenter le cours de Microsoft.

Les secteurs de l'énergie et de la finance sont des variables cycliques qui sont sensible aux variations macro-économiques (comme l'inflation).

Comparaison du prix actuel et des prédictions, à l'aide du Random Forest, pour le jour suivant et pour le mois suivant des actions ExxonMobil, Microsoft et JP Morgan



#### 4. Conclusion

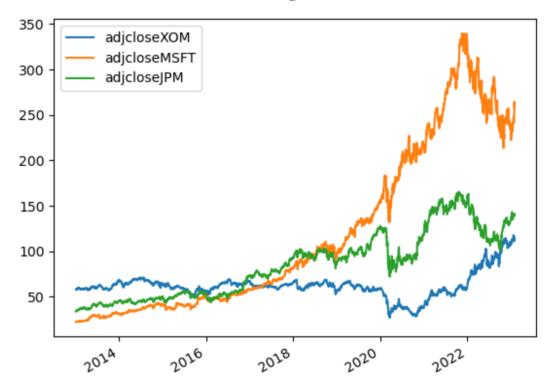
Nous avons comparé les prédictions des cours de l'action de 3 entreprises américaines, Exxon Mobil, Microsoft et JP Morgan, provenant de 3 secteurs différents, énergie, manufacturing/technology et finance/banque, obtenus à l'aide de 2 méthodes de machine Learning: LSTM (sous-catégorie du Recurrent Neural Network) et Random Forest. Le LSTM fait des prédictions sur la série temporelle alors que le Random Forest est davantage cross-sectionnel. Les résultats sont très satisfaisants dans l'ensemble avec des erreurs moyennes absolues et relatives (en pourcentage) qui restent faibles dans l'ensemble, surtout pour le Random Forest. Les prédictions du LSTM sont plus fiables pour les entreprises de secteurs cycliques comme Exxon Mobil dans l'énergie et JP Morgan dans la finance. En effet, l'historique du cours de leur action semble contenir des schémas (patterns) que le LSTM arrive bien à reproduire et prédire en termes de série séquentielle. Pour Microsoft, le LSTM prédit moins bien le cours de l'action que les données réelles; le MSTM semble moins bien intégrer les variations inattendues (surtout envolée du cours au dernier trimestre 2021).

### 5. Annexes

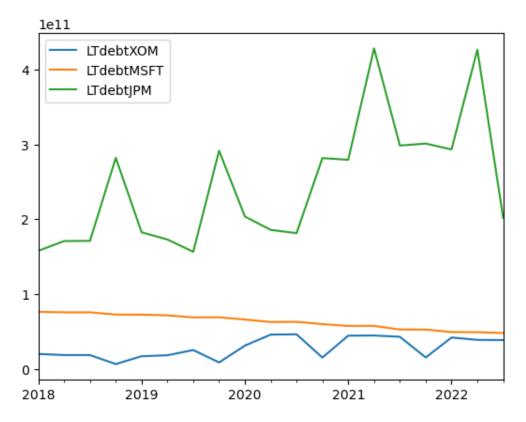
Statistiques descriptives

	adjcloseXOM	adjcloseMSFT	adjcloseJPM	CPImonthly	LTdebtXOM	LTdebtMSFT	LTdebtJPM	MeanSentXOM	MeanSentMSFT	MeanSentJPM
count	2540.00	2540.00	2540.00	120.00	2.000000e+01	2.000000e+01	2.000000e+01	121.00	121.00	121.00
mean	61.99	118.45	84.55	252.10	2.808065e+10	6.300440e+10	2.476209e+11	20.10	16.25	39.28
std	13.98	91.20	35.77	17.86	1.415468e+10	1.012696e+10	8.214416e+10	4.23	2.16	2.95
min	26.77	21.82	33.75	230.28	7.313000e+09	4.811600e+10	1.568580e+11	11.33	12.67	34.00
25%	57.68	41.51	49.53	237.89	1.731350e+10	5.328050e+10	1.795690e+11	17.33	14.67	37.33
50%	61.11	83.15	85.34	247.34	2.336550e+10	6.343950e+10	2.416265e+11	20.00	15.67	39.00
75%	64.81	199.95	108.98	258.78	4.289800e+10	7.240850e+10	2.919220e+11	23.00	18.00	40.33
max	117.76	339.89	165.25	298.01	4.688800e+10	7.690000e+10	4.277480e+11	33.67	21.33	53.67

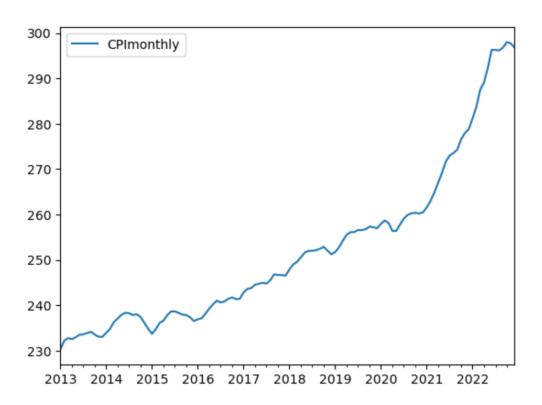
Évolution des cours ajustés de clôture journalière de l'action ExxonMobil, Microsoft et <u>JPMorgan</u>

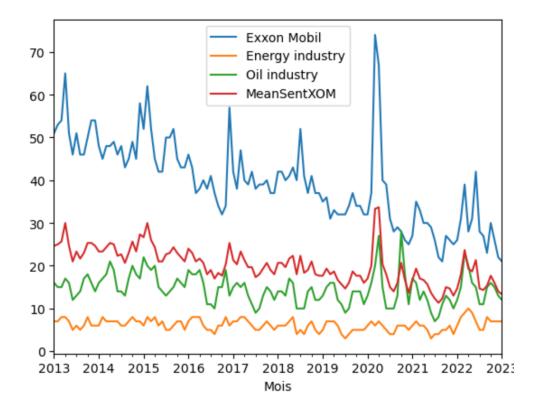


#### Évolution de la dette long terme de ExxonMobil, Microsoft et JP Morgan

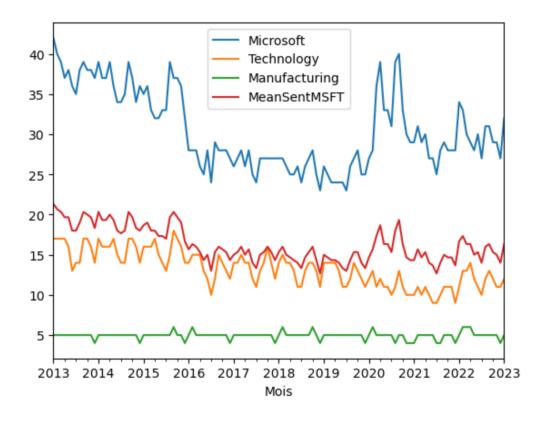


Évolution de l'inflation (CPI) mensuelle





Évolution des tendances et du sentiment moyen pour Microsoft



### Évolution des tendances et du sentiment moyen pour JP Morgan

