Projet 8 - Participez à une compétition Kaggle

G-Research Crypto Forecasting

Utiliser vos compétences en machine learning pour prédire le cours des crypto-monnaies

Cécile Guillot

Ingénieur Machine Learning

Introduction

Présentation de la plateforme Kaggle

Kaggle est un site communautaire proposant des compétitions et challenges en lien avec les data sciences et le machine learning. Ces compétitions sont proposées par des professionnels qui cherchent à avoir des idées innovantes concernant leur projet. Elles sont souvent accompagnées d'un cash prize. On y trouve aussi des datasets libres d'utilisation pour des projets personnels.

L'aspect communautaire permet de partager des travaux sous forme de notebook pour obtenir des retours mais aussi pour voir la manière dont les autres membres travaillent sur certains problèmes.

Présentation du projet

La compétition choisie ici a débuté le 2 novembre 2021 et se clôture le 2 février 2022. Elle a été proposée par le cabinet G-Research à Londres. Le but est de prédire les valeurs en dollars de crypto-monnaies.

Définitions des concepts

Le concept de crypto-monnaies a été introduit en 2008 avec la création du Bitcoin. Depuis 2015, ce concept s'est élargi et a permis la création de plusieurs devises se basant sur la technologie de la "blockchain". Les crypto-monnaies ont gagné un tel intérêt au fil des années que les marchés financiers s'y intéressent comme s'il s'agissait d'une devise physique. Il devient donc possible de spéculer sur les valeurs des crypto-monnaies pour investir dessus comme on le ferait avec des actions boursières.

Tout comme les données boursières, les crypto-monnaies sont ce que l'on appelle des séries temporelles. C'est-à-dire qu'il s'agit de données qui vont varier en fonction du temps. Les données financières se composent de plusieurs séries temporelles comme la valeur à l'ouverture d'une période de temps donné (minute, heure, jour), la valeur à la fermeture de cette période de temps, la valeur la plus élevée et la valeur la plus basse toujours dans une période de temps définie au préalable.

Les recherches sur la prédiction de données financières s'intéressent à trouver des algorithmes capables de prédire ces variables. Il existe des méthodes adaptées à ce type de données comme les modèles autorégressifs (ARMA, SARIMA, GARCH) ou les modèles de lissage exponentiel (Holt-Winters). Dans la littérature, on trouve aussi des méthodes transformant les problèmes de série temporelle en problème de régression.

Pour cela, une transformation des variables est effectuée pour pouvoir intégrer cette notion de temporalité à l'intérieur. Une fois ces transformations effectuées, on va pouvoir utiliser des algorithmes de machine learning de régression classiques. Cette méthode permet d'obtenir des résultats intéressants sur les prédictions de données de fermeture de la période temporelle d'intérêt.

Le but de ce projet est donc de montrer qu'il est possible de modifier des variables de série temporelles pour en faire des variables classiques utilisables dans un modèle de régression. Dans un second temps, des méthodes de machine learning classiques seront utilisées pour montrer qu'elles peuvent s'appliquer à des données temporelles après les avoir transformées.

Méthodologie

Recherche initiale

La méthode utilisée ici est une adaptation d'une recherche sur la prédiction de la valeur d'actions boursières de grande compagnie (Pfizer, Goldman-Sachs, *etc.*) réalisée en 2020 par Vijh et al.

Outils

Les analyses ont été faites sur l'outil Google Collaboratory. Ce choix a été fait pour pouvoir bénéficier d'un GPU lors de la phase de modélisation à l'aide de réseau de neurones.

Deux notebooks Jupyter ont été produits à l'issue de ce travail : un notebook "Analyses" et un notebook "Modélisation".

Pour réaliser cette étude, le langage de programmation utilisé est Python. Les libraries de data science suivantes ont été utilisées : Dask, Finta, Pandas, Matplotlib, Numpy, Pandas, Plotly, Scikit-Learn et Tensorflow.

Données

Les données fournies pour ce challenge sont téléchargeables sur la plateforme Kaggle : https://www.kaggle.com/c/g-research-crypto-forecasting. Le jeu de données utilisé pour l'entraînement comprend plusieurs colonnes mais seulement certaines vont être utilisées dans la suite de notre étude.

- **Timestamp** : horodatage de la donnée en format Posix
- **Asset_ID** : identifier la crypto-monnaie
- **Open** : valeur de la crypto-monnaie à l'ouverture de la minute
- **Close** : valeur de la crypto-monnaie à la fermeture de la minute
- **High** : valeur la plus élevée en dollar de la crypto-monnaie au cours de la minute
- **Low** : valeur la plus basse en dollar de la crypto-monnaie au cours de la minute
- **Volume** : volume de crypto-monnaie échangé au cours de la minute
- **Target** : valeur à prédire pour chaque minute

Les données couvrent une période allant du 1er janvier 2018 au 21 septembre 2021. Les enregistrements sont réalisés minute par minute. Lorsque l'enregistrement n'a pas eu lieu, cela a donné une non-existence de la ligne en question. On n'a donc pas de données manquantes sur ce point.

Cependant l'observation des valeurs manquantes a montré qu'il y en avait environ 3% dans la colonne de la variable "Target". Ces données ont donc été supprimées. Le calcul de la variable "Target" peut être trouvé sur le site Kaggle.

Les données à notre disposition portent sur 14 crypto-monnaies différentes : Binance Coin, Bitcoin, Bitcoin Cash, Cardano, Dogecoin, Ethereum, Ethereum Classic, IOS.IO, IOTA, Litecoin, Maker, Monero, Stellar et TRON. La plus ancienne est le Bitcoin. On va retrouver des monnaies plus récentes. Ces différences de création mais aussi d'algorithmes de cryptage (SHA256-d pour le Bitcoin ; Scrypt pour Litecoin par exemple) sur laquelle elles se reposent va expliquer des différences de valeurs propres en dollar.

Features engineering

Avant de réaliser cette étape, toutes les variables ont été transformées par l'intermédiaire d'une transformation logarithmique. Cette transformation permet de rendre les séries temporelles stationnaires. Il est aussi possible d'y ajouter une étape de différenciation.

Pour des raisons pratiques, le choix a été fait de réaliser des prévisions à l'échelle de l'heure.

Variables selon Vijh et al. (2020)

A l'aide des variables typiques de l'analyse financière (open, close, high et low), on va pouvoir construire de nouvelles variables. On aura deux types de variables : le premier type qui sera le résultat de différences et le second type qui va nous permettre d'incorporer la donnée temporelle dans la variable.

Les nouvelles variables construites vont donc remplacer les données déjà présentes. Cette précaution est prise pour éviter la fuite de données et l'utilisation d'information en double.

Les différences

Deux variables vont être créées à partir de différences. La différence entre la valeur la plus haute et la plus basse (H-L) et la différence entre la valeur d'ouverture et de fermeture (O-C) vont être calculées.

Les variables intégrant la donnée temporelle

L'utilisation de moyenne mobile va permettre d'intégrer la temporalité dans nos données. On va donc calculer les moyennes glissantes sur les valeurs à la fermeture. Les fenêtres glissantes choisies sont de 7, 14 et 21 jours. L'écart-type glissant sur 7 jours a aussi été calculé. L'inconvénient de cette méthode est l'introduction de valeurs manquantes. Ces valeurs manquantes sont donc remplacées par la valeur 0 dans notre jeu de données.

L'utilisation des moyennes glissantes permet donc d'incorporer la notion de temporalité dans nos données. On peut donc se passer de la présence de l'horodatage pour continuer notre étude.

Ajout de variables financières

Des données en lien avec le domaine des finances ont été ajoutées à notre jeu de données. Il s'agit de la moyenne mobile adaptative fractale (Fractal Adaptative Moving Average - FRAMA) et l'indice de force relative (Relative Strength Index - RSI).

Ces deux variables ne sont initialement pas présentes dans l'article de référence.

FRactal Adaptive Moving Average (FRAMA)

La moyenne mobile adaptative fractale est un indice utilisé dans le domaine de la finance. Il permet de mieux suivre les modifications de cours. Quand les modifications vont être importantes, l'indice aura une valeur importante. En revanche, quand le cours aura tendance à stagner, la valeur sera faible. Elle se calcule en utilisant les valeurs open, close, high et low.

Relative Strength Index (RSI)

L'indice de force relative permet d'avoir une information sur l'oscillation du momentum. Cet indice oscille entre 0 et 100. Une valeur supérieure à 70 indique que la devise est en surplus d'achats. Une valeur inférieure à 30 indique que la devise est en sous-achat. Cet indice permet aussi de savoir s'il va y avoir l'apparition d'une bulle spéculative.

Métrique d'évaluation

Pour évaluer les performances d'un modèle de régression, il existe plusieurs métriques comme le R-squared, le Mean Squared Error, le Root Mean Squared Error et le Mean Absolute Percentage Error.

La comparaison des performances sera faite à l'aide du Mean Absolute Percentage Error. Cet indice évalue le pourcentage en valeur de l'erreur moyenne. Ainsi, on peut obtenir une fourchette indiquant le pourcentage d'erreurs dans les prédictions de notre algorithme. De plus, le fait d'avoir un indice s'exprimant en pourcentage permet d'avoir une idée de l'ordre de grandeur plus simplement que si on utilisait une métrique qui se base sur l'unité de notre variable d'intérêt comme ça serait le cas avec le Mean Squared Error ou le Root Mean Squared Error.

Présentation des modèles utilisées

Initialement dans l'article de Vijh et al. (2020), on retrouve uniquement deux modèles : le Random Forest et un réseau de neurones. En plus de ces méthodes, la régression linéaire et le GRU ont été ajoutés à la comparaison.

Régression Linéaire

La régression linéaire est l'un des modèles classiques de machine learning lorsque l'on cherche à résoudre un problème de régression. Sa notation vectorielle est la suivante :

$$y_i = x_i'\beta + \varepsilon_i$$

Ce modèle est utilisé ici pour avoir un modèle de base. On va donc chercher à voir si les autres modèles utilisés auront des performances meilleures que la régression linéaire qui reste un modèle assez simple.

Random Forest

L'algorithme de random forest (ou de forêt aléatoire en français) est un algorithme qui fait partie des méthodes ensemblistes. On va entraîner de multiples arbres de décisions qui vont permettre d'obtenir une prédiction beaucoup plus proche de la réalité. Cet algorithme se base sur le principe de "wisdom of crowd", c'est-à-dire que la performance de plusieurs mauvais apprenants va être meilleure que la performance d'un bon apprenant seul. A la fin de l'exécution de l'entraînement de nos arbres, une moyenne est calculée pour avoir la prédiction de nos différents arbres.

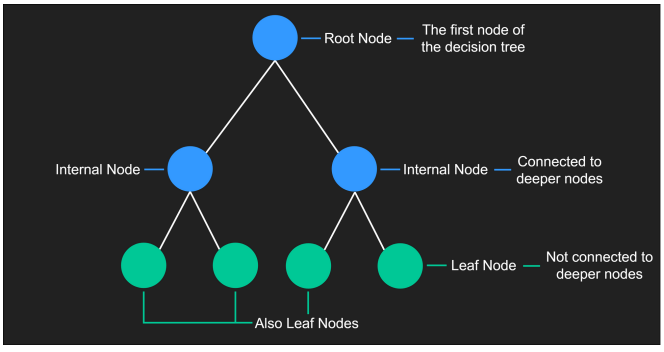


Illustration du fonctionnement d'un arbre de décision (source : https://mlfromscratch.com/decision-tree-classification/#/)

L'inconvénient de ce type de modèle est qu'il peut très vite faire du sur-apprentissage et ne pas se généraliser à des données nouvelles.

Réseau de neurones

Les réseaux de neurones se basent sur la formalisation du neurones biologiques en neurones virtuels. On va regrouper ces neurones ensemble pour former des couches et l'addition de plusieurs couches va former un réseau de neurones.

L'architecture choisie pour ce réseau se base sur celle qu'on retrouve dans l'article de Vidha et al. (2020).

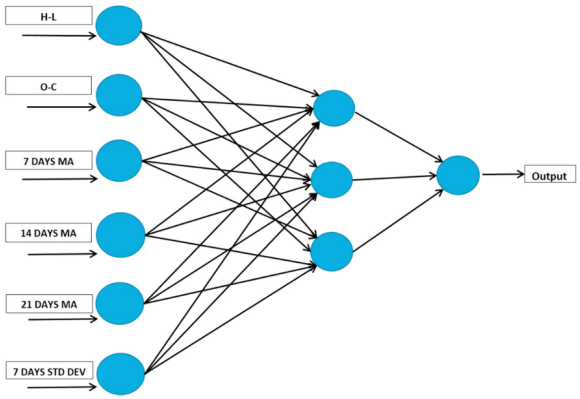


Illustration du réseau de neurones utilisés (tiré de Vijh et al., 2020)

Gated Recurrent Unit

Les modèles récurrents permettent d'intégrer une notion de temporalité ou de séquences dans les données que l'on cherche à modéliser dans une optique de prédiction. Parmi les nouveaux modèles de réseau de neurones récurrents, on trouve le LSTM (Long Short-Term Memory) et une de ces variations le GRU (Gated Recurrent Unit). Dans ces modèles, on retrouve des couches de neurones qui sont capables de se souvenir de ce qu'ils ont vu auparavant pour effectuer une prédiction sur une séquence. Les Gated Recurrent Unit permettent d'avoir la possibilité de retenir de l'information pour ensuite l'oublier quand cela est nécessaire.

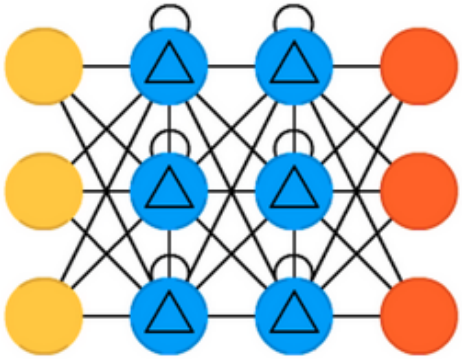


Illustration d'un réseau de neurones récurrent avec des GRU

Ce type de réseau de neurones est utilisé pour la reconnaissance de chaînes de caractères qui sont des informations séquentielles. On commence à les voir apparaître dans le domaine de la finance pour la prédiction de série temporelle.

Résultats

Analyses descriptives des données

Dans un premier temps, on s'est intéressé à l'analyse des données de crypto-monnaies de manière individuelle. Plusieurs datasets contenant les variables d'une seule crypto-monnaie ont été créés. A titre d'illustration, seuls les variables d'une seule crypto-monnaie sont présentés dans ce rapport. Les autres graphiques sont consultables dans le notebook 1.

La popularité, l'ancienneté et la technologie sur laquelle reposent une crypto-monnaie va faire varier sa valeur propre en dollar. Cependant, on remarque que les oscillations des monnaies sont plus ou moins corrélées.

Indice de tendance centrale

Les données moyennes au cours des années d'enregistrement sont présentées dans le tableau ci-dessous.

	Open	Close	Low	High
Binance Coin	77.37	77.47	77.26	77.37
Bitcoin	15611.97	15652.66	15578.89	15611.97
Bitcoin Cash	495.69	497.19	494.18	495.69
Cardano	0.35	0.360	0.35	0.35
Dogecoin	0.07	0.07	0.07	0.07
EOS.IO	4.89	4.908410	4.879162	4.892409
Ethereum	706.10	707.93	704.41	706.10
Ethereum Classic	15.39	15.50	15.32	15.39
IOTA	0.64	0.64	0.63	0.64
Litecoin	98.09	98.43	97.78	98.09
Maker	1899.80	1903.42	1896.445	1899.79
Monero	139.06	139.31	138.80	139.06
Stellar	0.17	0.17	0.17	0.17
TRON	0.03	0.035	0.03	0.03

Tableau des valeurs moyennes en dollar américain au cours des 3 dernières années pour 14 crypto-monnaies

Le Bitcoin est la devise ayant la valeur la plus forte parmi toutes nos crypto-monnaies. On remarque que les autres devises n'arrivent pas à atteindre de telles valeurs. De plus, on remarque que certaines monnaies ont des valeurs inférieures au dollar symbolique.

Evolution du cours au cours des deux dernières heures d'enregistrement

Pour avoir un aperçu du comportement des cryptomonnaies, l'évolution du cours de la monnaie sur les 120 dernières minutes d'enregistrement a été représenté par

l'intermédiaire d'un diagramme en chandeliers. Pour chaque minute, une chandelle va donner des informations sur la valeur à l'ouverture, à la fermeture, la valeur la plus élevée et la plus faible.

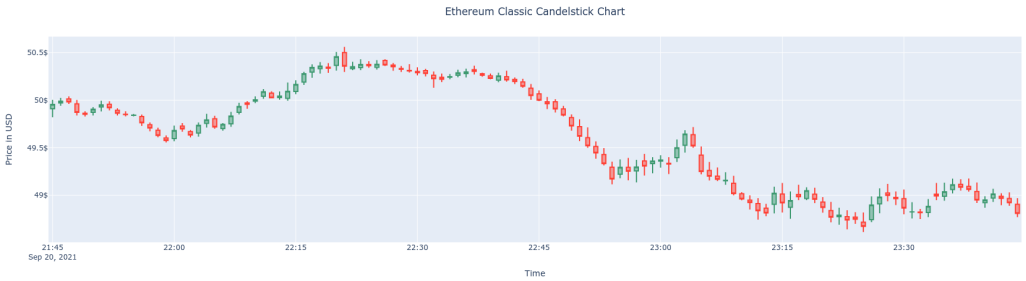
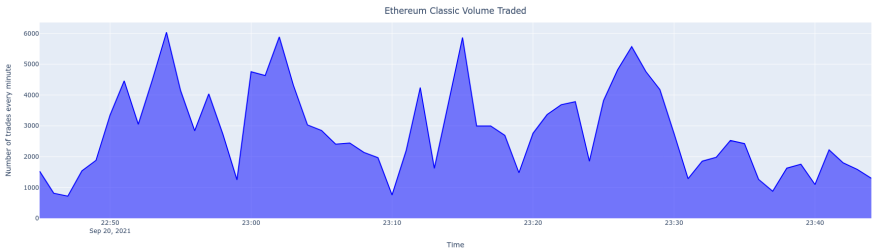


Diagramme en chandeliers de l'Ethereum classic durant 120 minutes

La valeur de l'Ethereum Classic varie entre $49 et $50.5 au cours des 120 dernières minutes de notre enregistrement. On voit que la valeur chute brusquement en une dizaine de minutes (entre 22:44 et 22:52). Cette diminution est aussi visible sur les autres crypto-monnaies au même instant. L'Ethereum Classic fait partie des devises les plus stables comparées à d'autres devises où les données d'ouverture, de fermeture, de valeur la plus haute et la plus basse vont se distribuer sur une échelle de valeur plus importante. Ce sera le cas d'IOTA mais il faut garder à l'esprit que la valeur de cette monnaie est basse (aux alentours d'un dollar) et que cela peut expliquer cette grande variabilité.

Volume échangé au cours de la dernière heure

Une autre variable observée concerne le volume échangé au cours de la dernière heure d'enregistrement. Il faut garder à l'esprit que les crypto-monnaies comme le Bitcoin sont disponibles de manière limitée. Il ne s'agit pas d'une ressource qui peut se créer indéfiniment. Cependant les volumes échangés donnent une information sur la popularité de la devise.

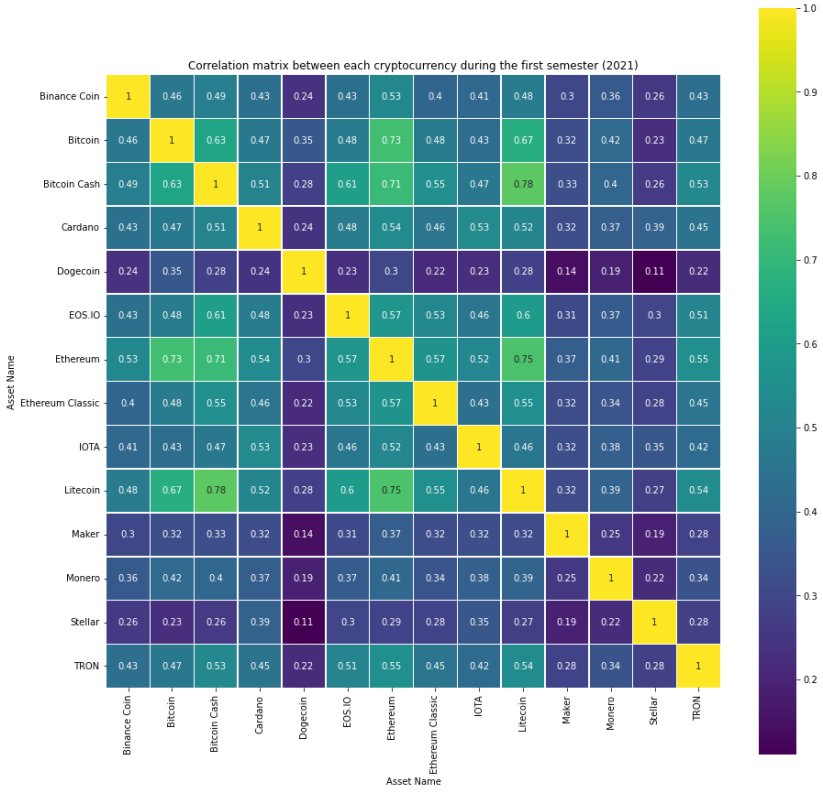


Graphique représentant le volume d'Ethereum Classic échangé en une heure

Le nombre de transaction effectué à la même morphologie pour les différentes devises. En revanche, la valeur propre d'unités échangées est différente selon la monnaie. On remarque que lorsque la valeur de la monnaie diminue, le volume des échanges diminue aussi avant de ré-augmenter.

Corrélation entre les différentes devises (1er semestre 2021)

Pour comprendre les liens entre les différentes crypto-monnaies, les coefficients de corrélation entre les monnaies ont été observées à l'aide d'une matrice de corrélation. Pour rendre la lecture plus simple, la matrice de corrélation a été effectuée sur 6 mois.



Matrice de corrélation entre les différentes crypto-monnaies au cours du 1er semestre 2021

Toutes nos devises ont des corrélations moyennes voire fortes entre elles. Seuls cinq monnaies semblent avoir une évolution décorrélée de la monnaie de référence (le Bitcoin). Il s'agit du Dogecoin, de Maker, de Monero, de Stellar et de TRON. On remarque que ces monnaies qui sont décorrélées du marché sont aussi celles qui vont avoir une valeur moyenne inférieure au dollar américain au cours des 3 dernières années.

Comparaison des performances des modèles

Une séparation des données en jeu d'entraînement et de test a été effectuée. Les jeux de données d'entraînement et de test étaient les mêmes pour tous les modèles.

Des jeux de validations ont été générés lors de l'entraînement des modèles pour les réseaux de neurones.

Les résultats obtenus pour le pourcentage d'erreur absolu sont présentés dans le tableau ci-dessous.

	Mean Absolute Percentage Error
Régression linéaire	12918659865.34%
Random Forest	9640191753.06%
Réseau de neurones	18.2%
Gated Recurrent Unit	13.3%

Tableau des résultats obtenus après entraînement de nos différents modèles

On observe que les performances pour les modèles de machine learning classiques sont très faibles. Ces modèles font d'énormes erreurs de prédiction qui les rendent inutilisables pour la tâche de régression que l'on cherche à résoudre.

En revanche, on remarque que les modèles basés sur les réseaux de neurones sont bien plus performants. Le pourcentage d'erreur moyenne reste important (entre 15 et 20%) mais ces modèles semblent plus exploitables que les modèles de Machine Learning.

Conclusion

Résumé

Le but de ce projet était de montrer s'il était possible d'appliquer des méthodes de machine learning utilisées dans le domaine de la prédiction des actions boursières pour des crypto-monnaies.

Les données collectées sur les crypto-monnaies nous permettent de faire un feature engineering semblables à celle que l'on pourrait effectuer pour des actions boursières. On dispose d'informations comme la valeur à l'ouverture, à la fermeture ainsi que les valeurs les plus hautes et les plus basses pour effectuer les différentes transformations. On peut donc conclure que la transformation de variables concernant les crypto-monnaies peut se faire de manière identique à celle que l'on ferait sur des actions boursières.

Dans une seconde partie, on a cherché à voir si les modèles utilisés de machine learning classique étaient applicables aux crypto-monnaies pour prédire une valeur cible dérivée de la valeur de fermeture.

Perspectives et limites

Bibliographie

Abadi, Mart';in, Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … others. (2016). Tensorflow: A system for large-scale machine learning. In *12th $USENIX$ Symposium on Operating Systems Design and Implementation ($OSDI$ 16)* (pp. 265–283).

Arte. (2021). Le mystère Satoshi : Aux origines du Bitcoin. https://www.youtube.com/watch?v=0ETcLj5jBy4

van Veen, F. & Leijnen, S. (2019) The Neural Network Zoo, https://www.asimovinstitute.org/neural-network-zoo/

Breiman. (2001). Random Forests, *Machine Learning*, 45(1), 5-32.

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Chung, Junyoung, et al. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555

Fernando, J. (2021). Relative Strength Index (RSI). https://www.investopedia.com/terms/r/rsi.asp

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &; Engineering*, *9*(3), 90–95.

Inc., P. T. (2015). Collaborative data science. Montreal, QC: Plotly Technologies Inc. Retrieved from https://plot.lys

Kaabar, S. (2021). Fractal Adaptive Moving Averages. The Full Guide. https://medium.com/the-investors-handbook/fractal-adaptive-moving-average-the-full-guide-c0ae348d9497

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, Gaël, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pironi, D. (2020). The Crypto Arbitrage Opportunity. https://suiteki.medium.com/the-crypto-arbitrage-opportunity-986b78740155

Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*.

Vijh, M., Chandolab, D., Tikkiwalb V. A., Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques, *Procedia Computer Science*, 167, 599-606

Wikipedia, Gated recurrent unit, https://en.wikipedia.org/wiki/Gated_recurrent_unit

Wikipedia, List of cryptocurrencies. https://en.wikipedia.org/wiki/List_of_cryptocurrencies

Annexe I : Liens vers les kernels Kaggle

EDA and Exponential Smoothing on Bitcoin :
https://www.kaggle.com/cecileguillot/eda-and-exponential-smoothing-on-bitcoin

Examples of Feature Engineering on 3 cryptocurrencies :
https://www.kaggle.com/cecileguillot/examples-of-feature-engineering-on-3-cryptos

Naive Drift And Regression Approaches :
https://www.kaggle.com/cecileguillot/naive-drift-and-regression-approaches

Annexe II : Kernels Kaggle utilisés pour ce projet

https://www.kaggle.com/thebrownviking20/intro-to-recurrent-neural-networks-lstm-gru

https://www.kaggle.com/odins0n/g-research-plots-eda