ANALYSE ET PRÉDICTION DE MARCHÉS FINANCIERS

Florian Rey

WEBTECH Lyon

1 Exploration et compréhension

1.1 Analyse exploratoire des données (Exploratory Data Analysis – EDA)

1.1.1 Importation, optimisation et nettoyage des données

Dans le but de garder des données correctement typées, j'ai converti chaque colonne dans le type qui me semblait le plus approprié, voilà le tableau de conversion :

Colonne	Type	Commentaire
Date	Datetime (format ISO8601)	Force le type date pour que ce soit reconnu par pandas comme tel.
Index	Category	Permet de décrire à pandas que les données sont catégorisables, le nombre d'index est limité, dans notre cas il n'y a que l'index du CAC40, mais il pourrait y en avoir d'autres dans le fichier cela serait compris
Open	Float32	Réduction de float64 vers float32 pour réduire l'utilisation mémoire
High	Float32	Réduction de float64 vers float32 pour réduire l'utilisation mémoire
Low	Float32	Réduction de float64 vers float32 pour réduire l'utilisation mémoire
Close	Float32	Réduction de float64 vers float32 pour réduire l'utilisation mémoire
Volume	Int32	Réduction de int64 vers int32 pour réduire l'utilisation mémoire

Afin d'optimiser les performances lors de l'import des données du CSV, j'ai fait le choix de transformer les données dans des types moins gourmands en mémoire, principalement les float64 et int64 vers float32 et int32.

Concernant la qualité des données du CSV, nous pouvons remarquer qu'il y a une quarantaine de lignes qui ne sont pas « utilisables », car les données ne sont pas valables soit parce qu'elles sont vides ou en erreur (ERR). Il y a également des valeurs qui sont potentiellement valides mais avec des symboles monétaires (€), j'ai donc fait le choix de les conserver mais en enlevant simplement le symbole de la valeur, afin que pandas puisse le convertir proprement en valeur numérique.

Concernant les dates et la cohérence temporelle, j'ai trié le jeu de données dans l'ordre croissant de date, pour permettre l'analyse des dates manquantes et la continuité. Grace à cela, j'ai pu voir qu'il manque dans les dates 237 valeurs, pour avoir une continuité parfaite de la date minimale à la date maximale, cela peut donc provoquer sur le graphique des sauts dans la courbe.

Pour prouver l'optimisation de la mémoire grâce au typage ainsi qu'au filtrage des données, j'utilise *df.info()* qui nous fait une analyse du jeu de données en mémoire, celui-ci nous retourne alors les informations suivantes :

Avant	Après

`<class 'pandas.core.frame.DataFrame'>`
`RangeIndex: 533 entries, 0 to 532`
`Data columns (total 7 columns):`
` #  Column  Non-Null Count  Dtype`
`---  ------  --------------  -----`
` 0  Date  533 non-null  datetime64[ns]`
` 1  Index  533 non-null  object`
` 2  Open  527 non-null  float64`
` 3  High  526 non-null  object`
` 4  Low  530 non-null  object`
` 5  Close  529 non-null  float64`
` 6  Volume  523 non-null  float64`
`dtypes: datetime64[ns](1), float64(3), object(3)`
`memory usage: 29.3+ KB`
	`<class 'pandas.core.frame.DataFrame'>`
	`Index: 494 entries, 0 to 532`
	`Data columns (total 7 columns):`
	` #  Column  Non-Null Count  Dtype`
	`---  ------  --------------  -----`
	` 0  Date  494 non-null  datetime64[ns]`
	` 1  Index  494 non-null  category`
	` 2  Open  494 non-null  float32`
	` 3  High  494 non-null  float32`
	` 4  Low  494 non-null  float32`
	` 5  Close  494 non-null  float32`
	` 6  Volume  494 non-null  int32`
	`dtypes: category(1), datetime64[ns](1), float32(4), int32(1)`
	`memory usage: 18.0 KB`

Ici la mémoire est à plus de 29,3 KB	Une fois optimisé nous sommes à 18 KB soit une économie de 11,3 KB, sur un petit dataset comme nous av

Nous pouvons voir l'évolution de la mémoire passant de 29,3 KB à 18 KB, ce qui représente une économie de 11,3 KB. Sur un petit jeu de données cela ne parait énorme, mais avec un plus gros jeu de données, cela pourrait être une différence plus importante pour la machine.

1.1.2 Analyse visuelle de base

Lors du nettoyage des données, j'ai pu récupérer la période sur laquelle je vais faire l'analyse visuelle, qui sera entre le 2 janvier 2023 et le 1er janvier 2025.

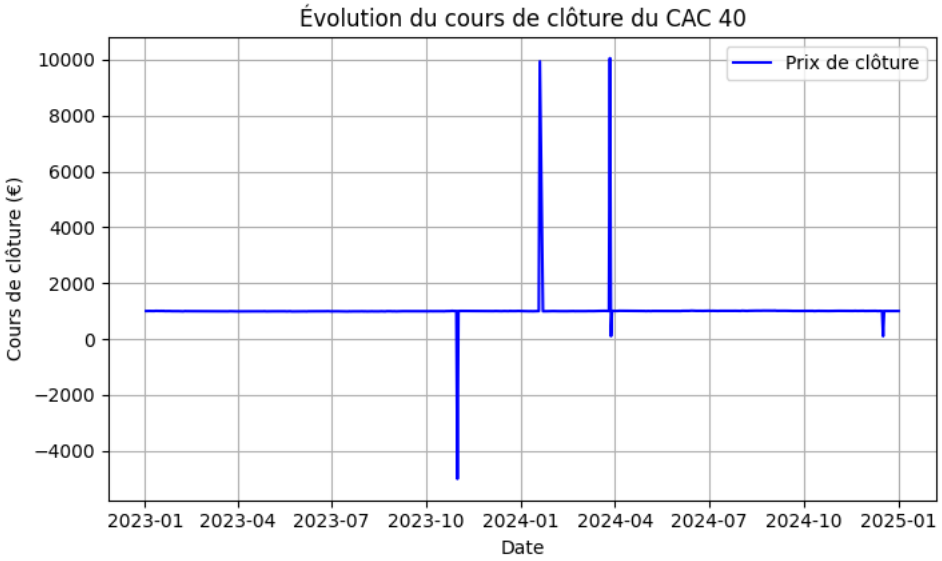

Figure 1 - Courbe de l'évolution du cours de clôture

Le graphique précédant représente l'évolution du cours de clôture de l'indice CAC40. On peut voir sur le graphique qu'il est majoritairement constant, hormis quelques exceptions qui semblent être des anomalies. Pour le filtre des anomalies, cela sera opéré dans la prochaine partie, mais il est intéressant de relever cela dès à présent.

Concernant la question de la corrélation entre les variables, prix, volume et volatilité, j'ai pu extraire un graphique visuel de la fonction *.corr()* de pandas, qui nous montre les corrélations entre chaque colonne :

Figure 2 - Corrélation entre les colonnes prix, volume et volatilité

On peut donc en conclure d'après cette visualisation qu'il n'y pas de flagrante corrélation entre ces trois colonnes. La volatilité étant toujours très faible et le prix de fermeture toujours identique, il est difficile de voir une corrélation entre les deux, les variations sont trop minimes ici.

Afin de visualiser la distribution des rendements, j'ai récupéré le pourcentage de rendement dans l'histogramme en limitant l'affichage entre -1 et 1. Les valeurs qui sont donc présentées sont en pourcentage, nous pouvons donc en conclure que le rendement du CAC40 se trouve majoritaire autour -1% et 1%.

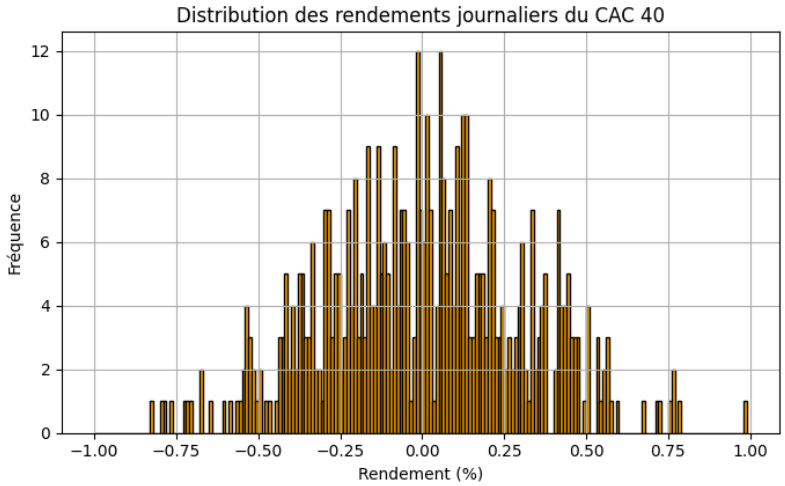

Figure 3 - Histogramme de distribution des rendements

2 Détection d'anomalies

Après avoir observé différentes courbes, on peut très nettement voir qu'il y a au moins quatre anomalies majeures potentielles.

- La première anomalie se trouve autour du 30 octobre 2023.
- La seconde anomalie se trouve autour du 18 janvier 2024.
- La troisième qui se trouve composée de 2 anomalies, une importante au 21 mars 2024 puis une plus petite au 28 mars 2024.
- Enfin la dernière anomalie se trouve autour du 16 décembre 2024.

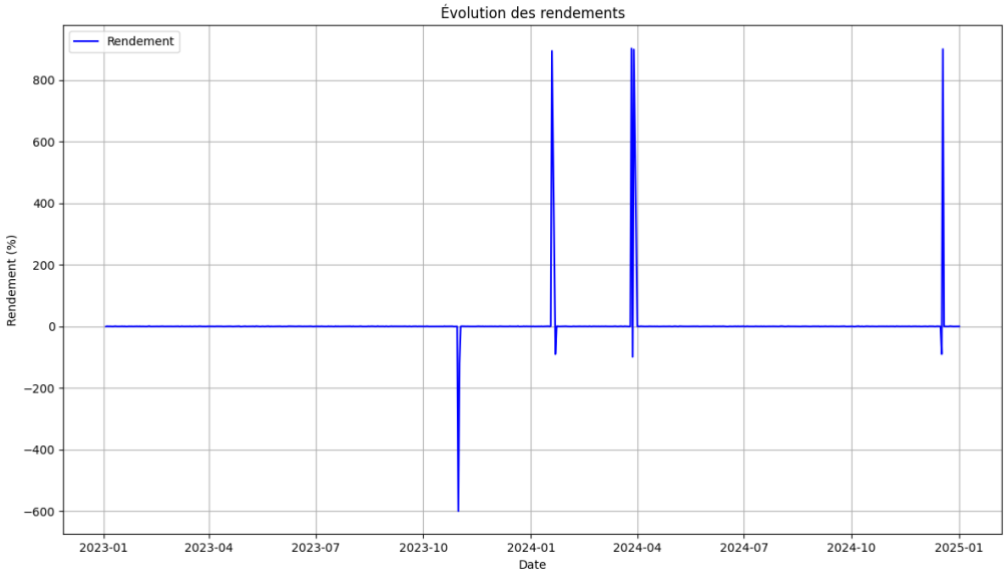

Figure 4 - Évolution des rendement journaliers

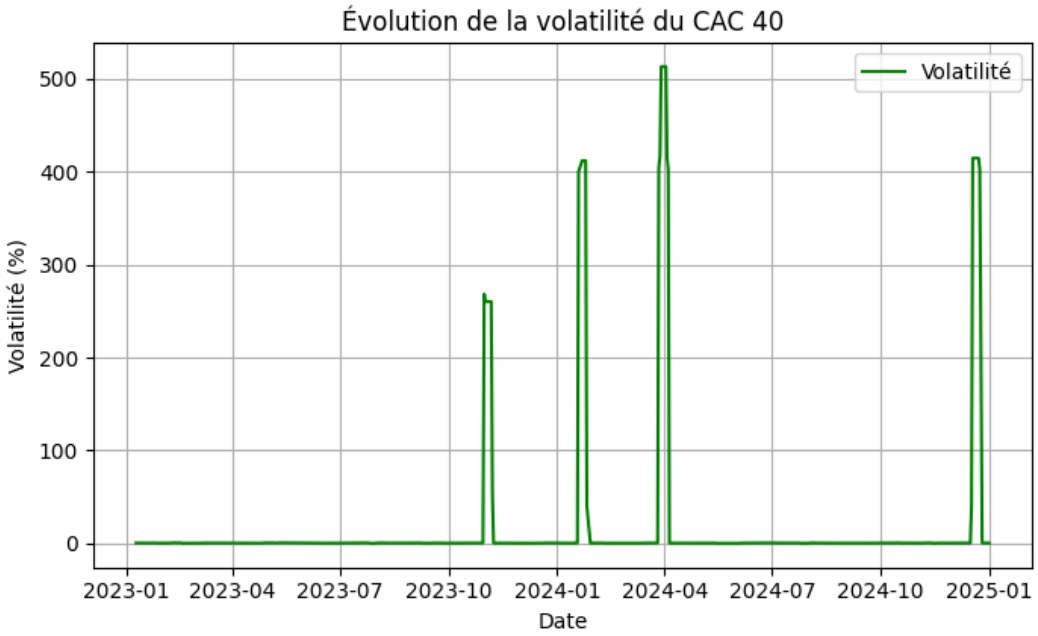

Figure 5 - Évolution de la volatilité

Les deux graphiques ci-dessus permettent d'étayer mes observations d'anomalie, notamment avec des ruptures de tendance très nettes, que ce soit sur l'évolution de la volatilité, ou l'évolution du rendement.

Afin de faire apparaitre les anomalies sur le graphique de l'évolution des cours de clôture, j'ai ajouté un script de détection des anomalies, pour pouvoir détecter et placer les pointeurs sur la courbe, voici le résultat :

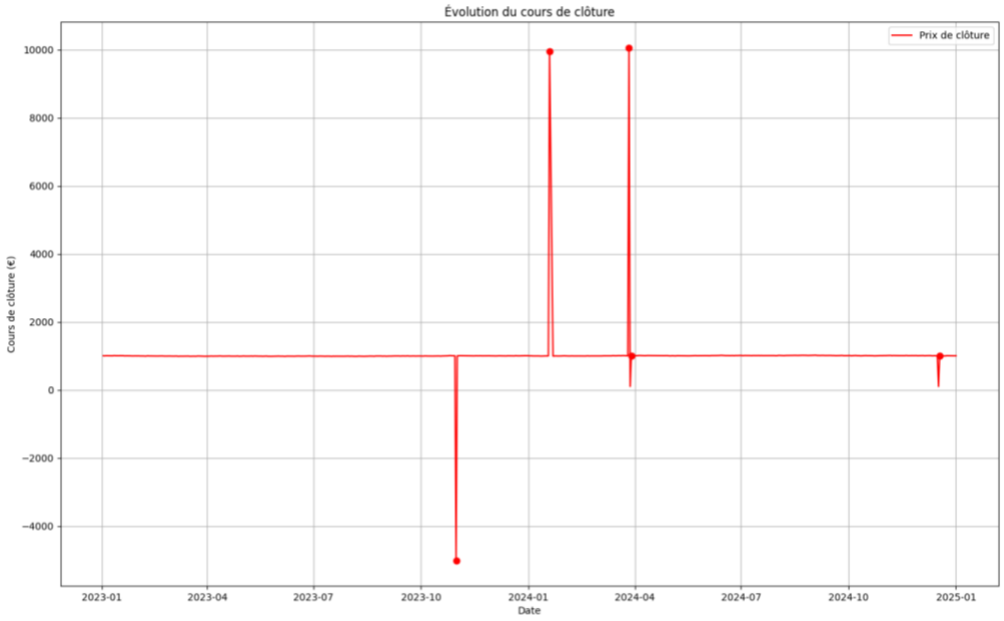

Figure 6 - Évolution du cours de clôture avec les anomalies

Sur les quatre anomalies relevées, deux pourraient être des anomalies expliquées, celle autour de janvier 2024 est possiblement explicable avec les mouvements des agriculteurs, une série de manifestation et de blocages routiers organisés. La seconde pouvant être expliquée est celle de décembre 2024, deux possibilités, il y avait déjà des tribunes de l'instabilité politique et des tensions budgétaires / coûts pour la France, mais il y a également il annonce de la banque de France concernant la fin d'un dispositif temporaire d'acceptation des créances privées supplémentaires (ACC) en date du 16 décembre.

3 Phase d'analyse statistique

Lors du développement du calcul de prédiction à la hausse ou à la baisse, j'ai testé plusieurs valeurs pour la comparaison des « quelques jours » précédant. Pour faire différents tests, j'ai décidé de tester sur les 9 jours précédant avec une boucle afin de déterminer quel est la plus grande précision que je pouvais atteindre. Et il en est ressorti que la meilleure valeur était 2 jours précédant en excluant la veille à 1 jours qui semblait être beaucoup trop proche et beaucoup trop haut par rapport aux autres.

```
Précision de la prédiction à 1 jour(s) : 71.26%
Précision de la prédiction à 2 jour(s) : 59.92%
Précision de la prédiction à 3 jour(s) : 58.30%
Précision de la prédiction à 4 jour(s) : 57.89%
Précision de la prédiction à 5 jour(s) : 57.89%
Précision de la prédiction à 6 jour(s) : 57.29%
Précision de la prédiction à 7 jour(s) : 58.30%
Précision de la prédiction à 8 jour(s) : 59.51%
Précision de la prédiction à 9 jour(s) : 57.69%
La meilleure prédiction est la prédiction à prediction_2 jour(s) avec une précision de 59.92%
```

On voit donc que finalement le modèle de prédiction à une précision à 59,92% soit presque 60%. Il y a donc à peu près deux chances sur trois que l'algorithme ait juste dans sa prédiction.

J'ai également fait une visualisation de la prédiction, que vous pourrez retrouver ci-après, les flèches vertes représentent les prédictions de hausse, et les flèches rouges représentent les baisses.

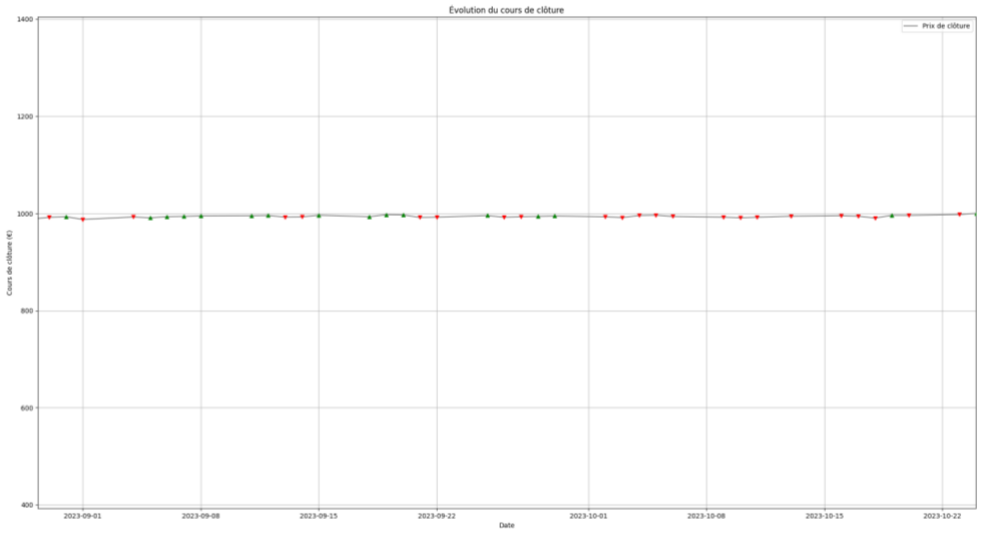

Figure 7 - Section de l'évolution du cours de clôture entre le 1er Sept. 2023 et le 22 Oct. 2023 avec les prédictions

4 Prédiction avec du Machine Learning

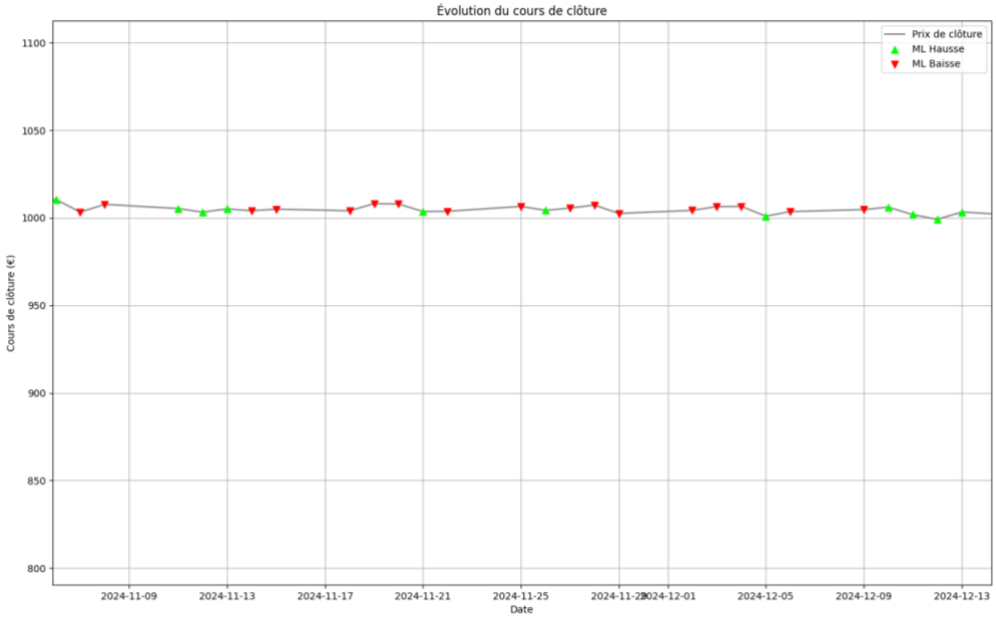

Figure 8 - Section de l'évolution du cours de clôture entre novembre 2024 et décembre 2024 avec les prédictions du par le ML

En utilisant les colonnes demandées, il est possible d'entrainer un modèle de Machine Learning, j'ai fait le choix de l'entrainer avec 80% des données et de le tester qu'avec

20%. En faisant cela j’obtiens le graphique ci-dessus, et les performances suivantes, 51,52% au tests de précision du modèle. Les performances sont donc un peu moins bonnes que l’algorithme précédant, car ici nous somme plus sur une chance sur deux que l’algorithme donne la bonne réponse.