Hackathon

QMI/LFIS/SESAMm

11-12 Mars 2022

Sujet 1 : Construction de portefeuille long-short robuste et performant

Le but du projet est de construire le portefeuille Long-Short avec la meilleure performance absolue sous contrainte de drawdown maximum.

Objectif : Trouver l'allocation avec le meilleur ratio Performance absolue divisée par Drawdown maximum sur un an roulant. Le drawdown maximum est la performance entre le point haut et le point bas sur l'historique d'une allocation.

Data : Les données disponibles sont les suivantes :

- données de marché : Des données observables comme le prix,le beta ou encore la volatilité réalisée
- données de sentiment : Des données exploitant les travaux de Sesamm sur le sentiment des articles traitant d'une société.
- données ESG : Des données exploitant les travaux de Sesamm sur l'aspect environnemental, social et gouvernance d'une société .
- données académiques : Des données exploitant les travaux de LFIS sur les différentes primes académiques

Vous pourrez soumettre vos prédictions durant la première phase de la compétition et recevoir automatiquement la performance absolue et le drawdown maximum de votre allocation ainsi que leur ratio (voir ci-dessous le processus d'évaluation).

Matériel : Durant la phase 1 vous disposerez des fichiers suivants :

- x_train_init.csv : les données avec le return associé sur la période d'entraînement.
- x_valid_init.csv : les données sans le return associé sur la période de validation.
- x_exemple_init.csv : Un exemple de poids à envoyer au bot.

Vous pourrez soumettre au bot des allocations sur la période de validation en nombre illimité. Pour la phase 1, chaque poids doit etre 1% ou 0% ou -1%. Ne peuvent être traitées pour un jour donné que les sociétés entrant dans la composition de l'indice.

Évaluation : Le fichier x_test_init.csv contient les données sur les observations de test. Vous ne pourrez pas recevoir votre score automatiquement sur ces échantillons durant la première phase de la compétition. Les organisateurs préviendront samedi matin les participants, sur le chat Telegram du Hackathon, pour annoncer la fin de la phase de validation et le début de la phase de test. Vous pourrez alors soumettre vos prédictions sur les échantillons de test, mais ne recevrez pas votre score. Il sera révélé à l'issue des soumissions de la totalité des équipes.

Phase finale : Après le déjeuner, les 4 meilleures équipes seront sélectionnées. L'épreuve concerne un autre univers géographique et les poids pourront varier entre 1% et 5% pour les poids positifs,-1% et -5% pour les poids négatifs. Les équipes disposeront alors des fichiers ci-dessous et ne pourront faire que 5 soumissions de validation. Un peu avant la fin de l'épreuve, elles auront deux tentatives pour contribuer sur le set de test et la dernière sera sélectionnée.

Matériel : Durant la phase finale vous disposerez des fichiers suivants :

- x_train_final.csv : les données avec le return associé sur la période d'entraînement.
- x_test_final.csv : les données sans le return associé sur la période de test.
- x_valid_final.csv : les données sans le return associé sur la période de validation.

Features :

Données de marché :

- **Date :** Un repère temporel correspondant à l'observation.
- **Sousjacent :** Une transco correspondant à un sous jacent donné.
- **secteurs :** Une transco correspondant au secteur.
- **Presence Indice** Un booleén indiquant si le sous jacent appartient à l'indice pour la date donnée.
- **vol :** La volatilité réalisée sur les 252 derniers jours
- **beta :** Le beta sur les 252 derniers jours.

Données académiques :

- **indic_mom_eu :** Un indicateur entre -1 et 1 définissant l'exposition à la prime momentum.
- **indic_qual_eu :** Un indicateur entre -1 et 1 définissant l'exposition à la prime qualité.
- **indic_val_eu :** Un indicateur entre -1 et 1 définissant l'exposition à la prime value.

Données de sentiment :

- **neutral_weighted_article :** La part de neutralité dans les articles traitant de la société.
- **neutral_weighted_entity :** La part de neutralité dans les phrases traitant de la société.
- **agreement_entity :** Un indicateur de la dispersion entre phrases positives et phrases négatives pour une société à l'échelon phrase.
- **agreement_average :** Un indicateur de la dispersion entre phrases positives et phrases négatives pour une société moyenné sur l'article et la phrase.
- **positive_weighted_article :** Un indicateur de la positivité moyenne à l'article pour une société donné.
- **positive_weighted_entity :** Un indicateur de la positivité moyenne à la phrase pour une société donné.
- **positive_index_article :** Un second indicateur de la positivité moyenne à l'article pour une socié donné.
- **positive_index_entity :** Un second indicateur de la positivité moyenne à la phrase pour une société donné.
- **volume_article :** Nombre d'articles pour le jour traitant d'une société donné.
- **volume_article_weighted :** volume_article weighté par la la similarité moyenne pour un jour donné pour une société donnée.

Données ESG :

- **volume_wordcount_environment_aggregated :** Nombre d'articles contenant au moins un keyword du topic environment aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_governance_aggregated :** Nombre d'articles contenant au moins un keyword du topic governance aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_risk_other_aggregated :** Nombre d'articles contenant au moins un keyword du topic risk_other aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_social_aggregated :** Nombre d'articles contenant au moins un keyword du topic social aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_all_risks :** Nombre d'articles contenant au moins un keyword du topic all_risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.

- **volume_wordcount_bankruptcy :** Nombre d'articles contenant au moins un keyword du topic bankruptcy moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_child_labor :** Nombre d'articles contenant au moins un keyword du topic child_labor moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_corruption :** Nombre d'articles contenant au moins un keyword du topic corruption moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_cybersecurity :** Nombre d'articles contenant au moins un keyword du topic cybersecurity moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_data_privacy :** Nombre d'articles contenant au moins un keyword du topic data_privacy moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_discrimination_racism_sexism :** Nombre d'articles contenant au moins un keyword du topic discrimination_racism_sexism moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_forced_labor :** Nombre d'articles contenant au moins un keyword du topic forced_labor moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_fraud_embezzlement_crime :** Nombre d'articles contenant au moins un keyword du topic fraud_embezzlement_crime moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_genocide :** Nombre d'articles contenant au moins un keyword du topic genocide moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_governance :** Nombre d'articles contenant au moins un keyword du topic governance moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_health_hazards :** Nombre d'articles contenant au moins un keyword du topic health_hazards moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_human_rights :** Nombre d'articles contenant au moins un keyword du topic human_rights moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_human_trafficking :** Nombre d'articles contenant au moins un keyword du topic human_trafficking moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_insider_trading :** Nombre d'articles contenant au moins un keyword du topic insider_trading moyenné par la similarité moyenne pour un jour donné pour une société donnée.

- **volume_wordcount_kyc :** Nombre d'articles contenant au moins un keyword du topic kyc moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_lawsuit :** Nombre d'articles contenant au moins un keyword du topic lawsuit moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_money_laundering :** Nombre d'articles contenant au moins un keyword du topic money_laundering moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_political_risk :** Nombre d'articles contenant au moins un keyword du topic political risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_regulatory_risk :** Nombre d'articles contenant au moins un keyword du topic regulatory risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_scandal :** Nombre d'articles contenant au moins un keyword du topic scandal moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_sexism :** Nombre d'articles contenant au moins un keyword du topic sexism moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_social :** Nombre d'articles contenant au moins un keyword du topic social moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_tax_avoidance :** Nombre d'articles contenant au moins un keyword du topic tax_avoidance moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **volume_wordcount_whistleblower :** Nombre d'articles contenant au moins un keyword du topic whisleblower moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_environment_aggregated :** Part négative des articles contenant au moins un keyword du topic environment aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_governance_aggregated :** Part négative des articles contenant au moins un keyword du topic governance aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_risk_other_aggregated :** Part négative des articles contenant au moins un keyword du topic risk_other aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_social_aggregated :** Part négative des articles contenant au moins un keyword du topic social aggrégé moyenné par la similarité moyenne pour un jour donné pour une société donnée.

- **negative_wordcount_all_risks :** Part négative des articles contenant au moins un keyword du topic all_risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_bankruptcy :** Part négative des articles contenant au moins un keyword du topic bankruptcy moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_child_labor :** Part négative des articles contenant au moins un keyword du topic child_labor moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_corruption :** Part négative des articles contenant au moins un keyword du topic corruption moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_cybersecurity :** Part négative des articles contenant au moins un keyword du topic cybersecurity moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_data_privacy :** Part négative des articles contenant au moins un keyword du topic data_privacy moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_discrimination_racism_sexism :** Part négative des articles contenant au moins un keyword du topic discrimination_racism_sexism moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_forced_labor :** Part négative des articles contenant au moins un keyword du topic forced_labor moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_fraud_embezzlement_crime :** Part négative des articles contenant au moins un keyword du topic fraud_embezzlement_crime moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_genocide :** Part négative des articles contenant au moins un keyword du topic genocide moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_governance :** Part négative des articles contenant au moins un keyword du topic governance moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_health_hazards :** Part négative des articles contenant au moins un keyword du topic health_hazards moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_human_rights :** Part négative des articles contenant au moins un keyword du topic human_rights moyenné par la similarité moyenne pour un jour donné pour une société donnée.

- **negative_wordcount_human_trafficking :** Part négative des articles contenant au moins un keyword du topic human_trafficking moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_insider_trading :** Part négative des articles contenant au moins un keyword du topic insider_trading moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_kyc :** Part négative des articles contenant au moins un keyword du topic kyc moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_lawsuit :** Part négative des articles contenant au moins un keyword du topic lawsuit moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_money_laundering :** Part négative des articles contenant au moins un keyword du topic money_laundering moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_political_risk :** Part négative des articles contenant au moins un keyword du topic political risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_regulatory_risk :** Part négative des articles contenant au moins un keyword du topic regulatory risk moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_scandal :** Part négative des articles contenant au moins un keyword du topic scandal moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_sexism :** Part négative des articles contenant au moins un keyword du topic sexism moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_social :** Part négative des articles contenant au moins un keyword du topic social moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_tax_avoidance :** Part négative des articles contenant au moins un keyword du topic tax_avoidance moyenné par la similarité moyenne pour un jour donné pour une société donnée.
- **negative_wordcount_whistleblower :** Part négative des articles contenant au moins un keyword du topic whisleblower moyenné par la similarité moyenne pour un jour donné pour une société donnée.