**Contexte**

Le marché américain des actions est l'un des plus liquides au monde, et fournit de nombreuses opportunités d'investissement. Les deux dernières heures de la session, entre 14H et 16H, sont les plus liquides de toutes.

Cette période est la préférée pour acheter ou vendre une grande quantité d'actifs, car les conditions de marché sont alors meilleures (coûts de transaction plus bas, volatilité moindre, …). C'est pourquoi estimer à l'avance le comportement d'un actif sur cette période permet d'optimiser l'ensemble de son portefeuille.

**But**

Le but de ce challenge est d'estimer la direction du prix d'une action durant les deux dernières heures de trading, connaissant le comportement du début de la journée.

Pour éviter le bruit habituel des mouvements financiers, nous ne considérons que trois directions :

* une baisse franche du prix ;
* une évolution mineure, quel que soit le sens ;
* une hausse franche du prix.

Description des données :

Nous fournissons les évolutions de prix (rendements) avec une granularité de 5 minutes, ce qui aboutit à 53 valeurs sur ces 4,5 heures, pour chaque jour et action.

Comme ces rendements sont faibles, nous les exprimons en point de base (bps), c'est-à-dire :

Nous donnons les colonnes :

* « Id » identifiant unique de cette ligne
* « day » identifiant du jour concerné concerné (non unique dans le jeu de données)
* « equity » : identifiant de l’action concernée (non unique dans le jeu de données)
* , le rendement sur les 5 premières minutes
* , le rendement sur les 5 premières minutes

Pour réduire la difficulté de la tâche, la prédiction se limite à estimer la direction du rendement des deux dernières heures, groupées en 3 cas :

* -1 si
* 0 si le
* +1 Si rendement > 25 bps

Le fichier de prédiction comprend les colonnes :

* **'ID'**, l'identifiant unique la ligne prédite ;
* **'reod'**, la classe estimée du rendement de la fin de session, incluse dans [−1,0,1] selon la formule susmentionnée.

Il est à noter que pour éviter toute fuite d'information, les **jeux d'entrainement** et de test ne comportent pas d'actions ni de jours en commun. Les actions des deux datasets ont cependant exactement les mêmes caractéristiques financières, afin de pouvoir prédire leurs comportements. Néanmoins, les deux périodes concernées sont totalement différentes, afin de correspondre à un vrai travail de prédiction, sur des données réelles d'un monde dont les caractéristiques peuvent changer.

**Description du benchmark**

* Comme ce travail de classification ne comporte que 3 classes, il est simple d'obtenir un score voisin de 33%33% avec des soumissions triviales, aléatoires, voire même fixes. C'est un moyen simple de tester vos soumissions.
* Le benchmark est moins naïf, et agrège les 53 rendements en caractéristiques plus haut niveau, puis les comparent entre elles en utilisant les données d'identifiant d'action et/ou de jour, afin de pouvoir détecter des modèles plus génériques. Un classifieur basique utilisant l'état de l'art actuel permet ensuite d'obtenir un score de 41,74%