Hackathon

QMI / LFIS / SESAMm

28-29 February 2020

Sujet : Prédiction de volatilité aux dates d'annonces de résultats

Le but du projet est de mettre en place une stratégie systématique qui profiterait du comportement de la volatilité (implicite et réalisée) des titres individuels autour des annonces de résultats des sociétés concernés.

Les investissements ne sont pas réalisés directement sur les sous-jacents mais sur leurs volatilités via des achats simultanés de call et de put (straddle).

L'achat de volatilité est une stratégie perdante sur le long terme (i.e. la volatilité implicite est en moyenne supérieure à la volatilité réalisée). Cependant, autour des dates d'annonce de résultats, les rendements présentent parfois des comportements plus volatiles que d'ordinaire.

Dans cet exercice, nous allons chercher à prédire la classe $Y = 1$, composée des échantillons pour lesquels l'amplitude du rendement de l'action, au jour de son annonce de résultat, fut plus grande qu'anticipée. Par opposition, la classe $Y = 0$ sera composée des échantillons pour lesquels l'amplitude du rendement de l'action fut plus faible que son anticipation.

Pour calibrer votre modèle, vous disposez de plus de 15000 observations. Chaque observation correspond à une date d'annonce de résultat. Ont été inclus dans ces observations tous les constituants du SP500 (américain, 500 stocks), et du Stoxx 600 (européen, 600 stocks), entre le début de l'année 2012 et la fin de l'année 2019.

Objectif Le but de ce sujet est de prédire le comportement volatile d'une action le jour de sa publication de résultats, à partir des observations suivantes:

- **sector**: un code représentant une classification sectorielle. 20 secteurs forment une partition de l'univers des stocks.
- **earnings_implied_obs**: l'anticipation de l'amplitude du rendement de l'action. Cette valeur est déduite des niveaux de prix des options (call et put) quelques jours avant la date de l'annonce de résultats.
- **delta_vol_1w (/1m/1y)**: l'évolution de la volatilité implicite à la monnaie, de maturité 3 mois, respectivement sur la dernière semaine, le dernier mois ou la dernière année.

- **return_1w (/1m/1y)**: Le rendement de l'action respectivement sur la dernière semaine, le dernier mois ou la dernière année.
- **implied_vol_3m**: le niveau de volatilité implicite, à la monnaie de maturité 3 mois, connu deux jours avant l'annonce de résultats.
- **realised_vol_1w (/1m/1y)**: l'écart-type des rendements de l'action respectivement sur la dernière semaine, le dernier mois ou la dernière année.
- **ratio_put_call**: Le ratio de l'open interest des puts et de la somme des open interest des puts et des calls. L'open interest est la quantité d'options (put ou call) en position ouverte.
- **publication_date_funda**: L'âge, en "jours" de publication des données fondamentales listées ci-dessous.
- **exchange**: `EU` si l'action appartient au Stoxx 600, `US` si elle appartient au SP500.
- **net_income**: Résultat net de l'entreprise, divisé par sa capitalisation boursière.
- **shareholders_equity**: Capitaux propres de l'entreprise, i.e. les ressources qui appartiennent à ses actionnaires, divisés par sa capitalisation boursière.
- **net_debt**: Dette nette de l'entreprise, divisée par la capitalisation boursière.
- **ebitda** : Bénéfices avant intérêts, taxes, dépréciations et amortissements, divisés par la capitalisation boursière.
- **ebit** : Bénéfices avant intérêts et taxes, divisés par la capitalisation boursière.
- **sales**: Chiffre d'affaire sur les 12 derniers mois, divisé par sa capitalisation boursière.
- **cash_flow**: Le flux de trésorerie de l'entreprise, divisé par sa capitalisation boursière.
- **payout_ratio**: Le payout ratio exprime le taux de distribution des bénéfices. Il est calculé en divisant le montant des dividendes distribués par les bénéfices nets consolidés de l'entreprise.
- **raw_id**: un identifiant technique. N'est d'aucune utilité pour améliorer la performance du modèle.

Matériel Durant la première phase de la compétition (phase 1), vous disposerez des fichiers suivants:

- `ref_train_x.csv`: les variables explicatives, sur les observations d'entraînement.

- `ref_train_y.csv`: les valeurs de la variable à expliquer, sur les observations d'entraînement. Ce fichier ne contient pas de header. Les lignes ne sont composées que d'une valeur 0 ou 1.
- `ref_test_x.csv`: les variables explicatives, sur les observations de test. Vous pourrez soumettre vos prédictions durant la première phase de la compétition et recevoir automatiquement le score obtenu (voir ci-dessous le processus d'évaluation).

Evaluation Le fichier `ref_valid_x.csv` contient les variables explicatives, sur les observations de validation. 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 (aux alentours de 10h), sur le chat Telegram du Hackathon, pour annoncer la fin de la phase de test, et de début de la phase de validation. Vous pourrez alors soumettre vos prédictions sur les échantillons de validation, mais ne recevrez pas votre score. Il sera révélé durant de déjeuner du samedi.

Votre score sera l'AUC de vos prédictions.

Il est ainsi attendu des participants, qu'ils soumettent des fichiers de prédiction qui respectent les normes suivantes:

- pas de header
- un float par ligne, correspondant à l'estimation de $P(Y = 1|X)$
- le point (".") comme séparateur décimal
- une ligne de moins que le fichier (valid-x) correspondant (car pas de header)