Applied big data analytics in finance

Master II FD

2024

TAC ECONOMICS

1 – Création de la base de données

- Code à faire sous python dans un fichier « 1_createDB.py »

- Cible à modéliser : choc sur une variable financière sur une fréquence quotidienne à un horizon de deux semaines.
  - Si un mouvement particulièrement fort à lieu dans 2 semaines, alors la cible sera à 1 sinon à 0.
  - Vous choisissez votre propre définition du choc (hausse, baisse, les deux, quel seuil ?, ...)
  - Ex : Indice boursier (S&P500, CAC40, ...) , taux de change (EURUSD, ...), prix du pétrole, commo (or, cuivre, ...), cryptos, et bien plus.

- Utiliser différents types de variables :
  - Indicateurs calculés de votre choix vous semblant pertinant
  - Indicateurs chartistes
  - Indicateurs de sentiment
  - Variables issues de méthodes type PCA
  - Variables macro, même si les fréquences ne sont pas quotidiennes

- Exportation de la base au format CSV une fois créée.

- Preférez l'utilisation du datalab, via vos APIKEY, mais vous pouvez utiliser tout ce que vous trouverez sur le web et jugez utile.

- Code à faire sous python dans un fichier « 2_modelisation.py »

- Utiliser n'importe quel(s) modèle(s)
  - Random Forest
  - Rpart
  - SVM
  - Regression logistique
  - XGBoost
  - Réseaux de Neurones
  - ...

- Ne pas utiliser une librairie qui test une baterie de modèles pour vous, je veux être sûr que vous connaissiez les modèles que vous utilisez dans cet exercice.

- Utiliser une base in-sample pour calibrer, et une autre out-of-sample pour tester les performances de vos modèles.
  - Ne pas utiliser d'échantillonnage aléatoire pour créer ces échantillons, mais des périodes continues
  - Si un modèle le permet, il est possible d'utiliser également un échantillon de validation

- Code à faire sous le même fichier « 2_modelisation.py » précédemment créé.

- Optimisez vos modèles en utilisant les metrics pertinantes de la matrice de confusion
  - Ne présenter que les résultats des individus out-of-sample.
  - Faites des boucles d'optimisation des paramètres que vous jugez important pour le(s) modèle(s) que vous utilisez

- Vous devez sortir une courbe ROC de vos prévisions ainsi que la matrice de confusion correspondant au seuil optimal que vous aurez définis.
  - Présentez également l'incide de Gini
  - Présentez les metrics que vous jugez pertinentes de la matrice de confusion obtenue avec ce seuil.

- Vous présenterez en 15 minutes par groupes votre projet en abordant les trois parties vues en cours ainsi que les difficultés rencontrées et les améliorations que vous souhaiteriez apporter au projet si vous aviez plus de temps.

- Sentez vous libre du format de cette présentation, l'idée est que je suis un responsable qui arrive dans le projet et qui veut comprendre les tenants et les aboutissants de votre travail !

- Je veux m'assurer durant cette présentation que :
  - Vous avez construit votre base intelligemment avec des indicateurs travaillés et non uniquement des données brutes.
  - Vous maitrisez le(s) modèle(s) que vous avez choisis d'utiliser
  - Vous comprenez très bien ce que sont les matrices de confusions et courbes ROC, et avez la connaissance des indicateurs de performance associés à cette matrice

Bon chance