Master IEF – Cours de ML
André GRONDIN
ag.m2dauphine@gmail.com
Projet

SUJET 1 - Création d'un petit automate

Pour chacune des questions suivantes, le script python, les résultats et éventuels remarques/interprétation devront être fournis. Les codes devront être commentés. Tous les travaux devront essentiellement être réalisés sous Python (Excel ou autre non toléré).
La construction d'une classe objet sera appréciée.

N'hésitez pas à prendre des hypothèses/initiatives intelligentes en cas de blocage ou de question ouverte.

Deux choix possibles pour votre rendu :

- un notebook alternant : énoncé de la question, code python commenté, résultats, remarques/difficultés rencontrées
- un script python commenté + un pdf avec résultats et remarques/difficultés rencontrées

Vous disposez d'une base de données comportant les caractéristiques de plusieurs actions.
Parmi ces caractéristiques, vous retrouverez :

- RDMT_x - rendement futur à 5 jours ouvrés
- HISTO_x - rendement historique de la valeur de l'indice à la clôture du marché
- VOL_x - écart historique des volumes d'échange
- UP_x - rendement historique de la valeur au plus haut en intraday
- DO_x - rendement historique de la valeur au plus bas en intraday

x = Suffixes J, S et M indiquant respectivement

- J = une période journalière (entre 2 jours ouvrés consécutifs)
- S = une période hebdomadaire
- M = une période mensuelle.

La base comporte également les mots parus au sein des actualités boursières journalières rattachées au ticker de la même ligne.

1. Donner les caractéristiques de la base :
   - nombre de lignes
   - nombre de colonnes
   - format/type des colonnes

2. Présenter quelques statistiques de la base. L'utilisation d'un graphique avec le package plotly sera appréciée.

3. Votre objectif : quel que soit le ticker, quelle que soit la variable binaire construite sur les rendements du ticker, obtenir le meilleur modèle suivant 2 des 3 métriques suivantes : auc, précision, recall.

   Il vous faudra construire la fonction prenant en input les features (variables explicatives), un variable à expliquer binaire. La fonction ne prendra en entrée que les lignes journalières d'un seul ticker.

   Pensez à écarter les variables aberrantes, peu volatiles, prospectives.

   Cette fonction doit comprendre les parties suivantes :
   - Normalisation
   - Binarisation
   - Retrait des variables trop corrélées (positivement ou négativement)
   - Choisir un des algorithmes suivants :
     - xgboost
     - GradientBoostingClassifier
     - neural_network.MLPClassifier
   - Analyser les hyperparamètres de l'algorithme retenu et assurer un grid computing de 10 n-uplets de paramètres
   - Une cross-validation de 4 folds est attendue
   - La fonction doit retourner
     - le n-uplet de paramètres
     - le résultat des 2 ou 3 métriques retenues
     - le top 4 de l'importance des variables

     où la maximisation de la performance et/ou la minimisation de l'erreur sont au rdv.

4. Vous bouclerez sur au moins 10 tickers de la base.

SUJET 2 - Importance des mots

Pour chacune des questions suivantes, le script python, les résultats et éventuels remarques/interprétation devront être fournis. Les codes devront être commentés. Tous les travaux devront essentiellement être réalisés sous Python (Excel ou autre non toléré).
La construction d'une classe objet sera appréciée.

N'hésitez pas à prendre des hypothèses/initiatives intelligentes en cas de blocage ou de question ouverte.

Deux choix possibles pour votre rendu :

- un notebook alternant : énoncé de la question, code python commenté, résultats, remarques/difficultés rencontrées
- un script python commenté + un pdf avec résultats et remarques/difficultés rencontrées

Vous disposez d'une base de données comportant les caractéristiques de plusieurs actions.
Parmi ces caractéristiques, vous retrouverez :

- RDMT_x - rendement futur à 5 jours ouvrés
- HISTO_x - rendement historique de la valeur de l'indice à la clôture du marché
- VOL_x - écart historique des volumes d'échange
- UP_x - rendement historique de la valeur au plus haut en intraday
- DO_x - rendement historique de la valeur au plus bas en intraday

x = Suffixes J, S et M indiquant respectivement

- J = une période journalière (entre 2 jours ouvrés consécutifs)
- S = une période hebdomadaire
- M = une période mensuelle.

La base comporte également les mots parus au sein des actualités boursières journalières rattachées au ticker de la même ligne.

1. Donner les caractéristiques de la base :
   - nombre de lignes
   - nombre de colonnes
   - format/type des colonnes

Nous allons zoomer sur les mots les plus fréquents des actualités boursières et créer un modèle de prédiction sur les rendements mensuels.

2. Présenter quelques statistiques de la base.

3. Montrer que, quel que soit le ticker, la liste des mots apparaissant au moins sur 400 lignes et dont le rendement mensuel est en moyenne supérieur à 1% est la suivante :

Mot	Nombre d'apparitions	Rendement mensuel moyen
part	716	0.0119
plus	418	0.0101
pour	1966	0.0136
euros	446	0.0104
group	493	0.0129
passe	527	0.0138
titre	467	0.0122
groupe	467	0.0146
nouvel	494	0.0115
releve	413	0.0103
actions	607	0.0102
capital	761	0.012
contrat	540	0.0106
nouveau	511	0.0119
capital.	751	0.011
nouvelle	415	0.012
objectif	636	0.0147
resultat	458	0.0112

Garder par la suite uniquement les lignes où au moins l'un des mots ci-dessus apparaît ; autrement dit que les variables explicatives portent pour modalité = 1.

4. Retrait des variables trop corrélées (positivement ou négativement).

Votre objectif est de construire un outil d'aide à la décision en cas d'apparition d'un des mots recensés ci-dessus et de recommander un acte d'achat ou de vente sur le ticker concerné.

5. Choisir un des algorithmes suivants :

- xgboost
- GradientBoostingClassifier
- neural_network.MLPClassifier

On laissera volontairement la présence de l'ensemble des tickers.

6. Analyser les hyperparamètres et assurer un grid computing d'environ 100 n-uplets de paramètres. Une cross-validation de 4 folds est souhaitée.

7. Définir la variable à expliquer comme une variable binaire sur les niveaux de rendements de la base filtrée et ce tout ticker confondu. Utiliser le package plotly pour dessiner un graphe de votre choix avec la variable à expliquer en ordonnée.

8. Obtenir le meilleur modèle suivant 2 des 3 métriques suivantes : auc, précision, recall ainsi que le top 4 de l'importance des variables.