**Exercice pratique No. 2**

**Date de remise : 14 avril 2020**

**Équipes : 3-4 personnes**

Le second travail pratique est une application des différentes techniques de classification montrées dans le cours.:

Données de base :

* Données sur des ETF.
  + Vous avez 50 ETF provenant de Blackrock de mars 2006 à maintenant
  + Données sont quotidiennes.
  + Objectif : Développez un système de trading qui sur-performe l’indice du SP500, soit l’ETF ‘’IVW’’.
  + Variables explicatives : Indicateurs techniques pour chaque ETF
    - Vous avez un programme Python qui vous donne 30 indicateurs techniques
    - En premier lieu, évaluez la performance de 5 indicateurs techniques les plus courants :
      * Bollinger Bands
      * Oscillateur stochastique
      * MACD
      * RSI
      * Money Flow
    - Pour faciliter l’utilisation des signaux techniques, il est indiqué de les transformer en catégories.
      * Le plus facile est d’avoir des signaux binaires par signal
        + Buy/Neutral
        + Sell/Neutral
      * Vous Pouvez également utiliser 3 catégories par signal technique :
        + Buy/Neutral/Sell
      * Vous choisissez le type de catégorisation
  + Vous calculez la performance de votre stratégie en utilisant le ‘’Adj Close’’.

L’objectif du TP2 est de comparer la performance des indicateurs techniques en utilisant différentes techniiques de classification

En premier comme dans toute analyse statistique, faites l’analyse descriptive de vos données.

1. Pour chacun des ETF, vous fournissez les statistiques de base
   1. Rendement moyen, volatilité, quantile, etc.
   2. L’analyse peut être présentée sous forme de tableaux et graphiques
2. Par la suite, vous validez la performance des 5 indicateurs techniques appliqués sur chaque ETF
   1. Vous calculez la performance du signal
      1. À chaque jour, vous allez avoir 3 situations
         1. ‘’Long’’ : Achat de l’ETF
         2. ‘’Neutral’’ : Vous maintenez votre position ou vous restez ‘’cash’’, c’est votre choix
         3. ‘’Short’’ : Vous prenez une position de vente à découvert
            1. Prenez le rendement inverse
      2. Ne tenez pas compte des frais de transaction ou du ‘’bid/ask’’ spread.
   2. Vous fournissez les statistiques sur la performance de la classification.
      * 1. En Python, vous utiliserez
           1. from Sklearn.metrics import accuracy\_score
      1. Matrice de confusion
         1. 

En Python, vous utiliserez

from sklearn.metrics import confusion\_matrix

* + 1. ROC Curve



En Python, vous utiliserez

From sklearn.metrics import auc

1. Développez un modèle qui combine l’ensemble des indicateurs techniques (soit les 5 ou l’ensemble des indicateurs, c’est vous qui décidez).
   1. En premier lieu, construisez un modèle unique pour tous les ETF. Ainsi, vous regroupez l’information sur les 50 ETF dans une seul fichier.
   2. Vous allez utiliser 5 types de modèles de classification
      1. Modèle Logit
      2. Arbre de décision
      3. Random Forest
      4. Gradient Boosting
      5. Support Vector Machine
         1. Classification linéaire
         2. Classification radiale
2. Vous séparez votre échantillon en deux portions
   1. Échantillon d’entraînement (70% du total)
   2. Échantillon de test (30% du total)
      1. from sklearn.model\_selection import train\_test\_split
      2. X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state = 0) # 0.3 test\_size means 30%
3. Débutez avec le modèle de base, soit le modèle Logit.
   1. Variable cible (1/0) – 1 si ça monte et 0 sinon
   2. Variables explicatives
      1. Signaux techniques (‘’Buy’’, ‘’Neutral’’, ‘’sell’’) ( Soit 5 ou 30)
   3. Le modèle estimé comporte un trop grand nombre de variables pour être utile et en plus il ‘’overfittera’’. Vous devez donc réduire le nombre de variables à seulement celles qui semblent avoir un impact important et enlever les autres. Voici deux méthodes souvent utilisées.
      1. La 1ère est d’utiliser l’indice AIC (Akaike).
         1. Vous partez de vos résultats et calculez la valeur de l’indice AIC.
         2. Vous enlevez certaines variables dont les paramètres sont non-significatifs, du mauvais signe ou semblent incohérents et vous recalculez la statistique AIC. Le modèle avec la valeur AIC la plus faible est toujours préféré.
         3. C’est une méthode de sélection non systématique mais qui vous aidera à comprendre votre modèle
         4. Pour les fins de l’exercice, présenter deux modèles de votre choix
      2. La seconde est d’utiliser l’élimination récursive.
         1. Méthode systématique où les variables sont éliminées progressivement en fonction d’un critère statistique, jusqu’à un modèle plus simple. Ce modèle est souvent appelé ‘’stepwise’’ ou ‘’feature selection’’.
         2. Vous trouverez ci-joint un code Python pour vous aider
            1. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>
         3. Présentez le modèle sélectionné. À cette étape, vous devez faire preuve de jugement si la solution ne fait pas de sens. Si c’est le cas, enlevez les variables problématiques et recommencez.
   4. Évaluez la capacité de classification.
      1. Pour l’échantillon d’entraînement et l’échantillon test.
      2. Présentez la matrice de confusion et le graphique ROC.
      3. Interprétez les résultats.
      4. Quelles sont les variables les plus importantes dans ce modèle?
4. Refaites l’exercice avec un arbre de décision
   1. En premier lieu, utilisez toutes les variables explicatives.
   2. Estimez votre modèle.
      1. Vous pouvez utiliser skearn – DecisionTreeClassifier
   3. Présentez la matrice de confusion et le graphique ROC.
      1. Pour l’échantillon d’entraînement et l’échantillon test
   4. Répétez l’exercices avec les variables sélectionnées par le modèle Logit
   5. Notez les principales différences et similitudes et expliquez les divergences.
5. Refaites l’exercice avec la méthode Random Forest.
   1. En premier lieu, utilisez toutes les variables explicatives.
   2. Estimez votre modèle.
      1. Vous pouvez utiliser sklearn.ensemble – RandomForestClassifier
   3. Présentez la matrice de confusion et le graphique ROC
      1. Pour l’échantillon d’entraînement et l’échantillon test
   4. Identifiez les variables les plus importantes (‘’Variable Importance’’)
      1. Présentez le graphique sur l’importance des variables.
   5. Comparez les variables sélectionnées par Random Forest au modèle Logit.
      1. Quelles sont les variables d’importance similaires.
      2. Quelles sont les variables différentes.
      3. Pourquoi les variables sélectionnées ne sont pas totalement similaires?
6. Répétez l’exercices pour des méthodes plus flexibles mais opaques.
   1. Répétez le point (4) en utilisant.
      1. Support Vector Machine
         1. Fonction linéaire
         2. Fonction radiale
      2. Si vous en avez encore le goût, vous pouvez utiliser le deep learning.
7. Comparer une dernière fois, l’ensemble des modèles, sélectionnez-en un et expliquez votre choix.
8. Répétez l’exercice mais cette fois-ci pour chaque ETF pris individuellement.
   1. Autrement dit, vous refaites l’exercice précédent mais pour chaque échantillon.
   2. Assurez vous d’avoir le même échantillon
      1. Créez des sous-échantillon de vos ‘’train and test sets’’
   3. Comparez la performance d’une estimation individuelle vs modèle regroupé de l’étape précédente.
   4. Pas besoin de présenter tous les résultats mais uniquement la performance soit le rendement du modèle agrégé vs modèles individuels.
9. Enfin, si vous voulez pousser encore plus loin (optionnel)
   1. Développez un modèle de trading pour l’ETF sur l’indice SP500, soit ‘’IVW’’.
   2. Utilisez comme signal l’ensemble des indicateurs techniques pour l’ensemble des ETF.
   3. Ajouter le rendement de la journée précédente de tous les ETF.
   4. Ajouter également le rendement de la semaine précédentes de tous les ETF.
   5. Utilisez un SVM (radial) ou un deep learning.
   6. Comparez votre performance avec le rendement du SP500, soit une position passive sur l’ETF ‘’IVW’’.
   7. Notez bien que cette section est optionnelle mais pourrait fort bien constituer un projet de session