**Exercice pratique No. 2**

**Date de remise : 23 avril 2019**

**Équipes : 2-3 personnes**

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

Données de base :

* Données sur des demandes de crédit dans une institution financière.
  + Variable cible : Demande acceptée (perçue comme un bon crédit) (1/0)
  + Variables explicatives : 20 variables
    - 7 continues
    - 13 sous forme de catégories
  + Transformation des variables de catégories.
    - La plupart des modèles statistiques ont de la difficulté à s’adapter aux variables qui comportent plusieurs choix. Dans ce cas, vous devez les transformer en choix binaire (dummy).
    - Lorsqu’il y a deux possibilités, un des possibilités prend la valeur 1 alors que l’autre prend la valeur de 0.
    - S’il y a plus d’un choix, vous transformez cette variable en plusieurs variables binaires selon la règle.
      * Ainsi, vous ne pouvez avoir un nombre équivalent de variables binaires au nombre de choix. Si cela arrivait, vous pourriez avoir une combinaison de plusieurs cas ayant fait des choix différents qui, combinés, donneraient une ligne de ```1``. Une telle situation rendrait l’inversion de matrice impossible ayant une corrélation parfaite avec la constante ou le biais.
      * Par exemple si vous avez deux choix, vous aurez donc une seule variable binaire. Si vous avez 3 choix, vous pouvez les regrouper en 2 choix. Soit le client opte pour le choix 1 et non, le choix 2 ou non. Vous pourriez regrouper également deux choix pour obtenir une seule variable binaire.
  + Vous trouverez le chiffrier Excel sur Moodle pour le cours.
    - Analyse\_du\_credit.xlsx
    - La base de données comprend l’information sur 1,000 clients.
    - Il s’agit d’une des bases de données traditionnelles utilisée pour tester des méthodes statistiques.
    - Cette base de données est intéressante actuellement car la plupart des institutions bancaires souhaitent raffermir leurs méthodes de sélection ou de rejet de crédit via l’utilisation de l’apprentissage machine.

L’objectif du TP2 est de comparer la performance de la classification de la variable ‘’credit’’ par différents modèles d’apprentissage statistique.

1. En premier comme dans toute analyse statistique, faites l’analyse descriptive de vos données.
   1. Mesures de base – Moy.,min, max, quantile, nombre etc.
   2. L’analyse peut être présentée sous forme de tableaux et graphiques
2. Par la suite, séparez votre échantillon en deux portions
   1. Échantillon d’entraînement (70% du total)
   2. Échantillon de test (30% du total)
3. Débutez avec le modèle de base, soit le modèle Logit.
   1. Variable cible (credit 0/1)
   2. Variables explicatives
      1. Dummies
      2. Variables continues
   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 (`).
         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. Sélectionnez les variables explicatives qui vous apparaissent les meilleurs pour expliquer le crédit.
7. Appliquez cet ensemble de variables sur les méthodes plus flexibles mais opaques.
   1. Répétez le point (4) en utilisant.
      1. Support Vector Machine
         1. Fonction linéaire
         2. Fonction radiale
      2. Modèle de neurones.
         1. Déterminez le nombre de couches cachées par tâtonnements (1,2,…)
      3. Si vous en avez encore le goût, vous pouvez utiliser le deep learning.
8. Comparer une dernière fois, l’ensemble des modèles , sélectionnez en un et expliquez votre choix.