**Projet de Scoring**
**M2 IMSD – Université Paris Saclay**
**Date limite : 15 Février**

# Evaluation Finale de Scoring

*Ce projet est un cas d'école mais appliqué à des vraies données d'entreprises (anonymes).*

*La qualité de la rédaction primera sur la quantité. Les démonstrations mathématiques doivent être étayées et convaincantes. De plus, les sorties « logiciel» doivent être toutes commentées.*

*Livrable attendu : fichier word (.doc) avec les codes SAS (ou R) en annexe.*

**Partie A** : régression logistique sur variables quantitatives

On souhaite modéliser le défaut d'un portefeuille composé de plusieurs entreprises afin d'anticiper le comportement d'entités nouvelles. Notons Y la variable défaut valant 1 si l'entreprise n'a pas fait défaut, 0 sinon. Vous disposez d'un ensemble de 5 variables quantitatives continues : WCTA, RETA, EBIT_TA, METL, STA.

1. Enumérez les différents modèles (paramétriques et non paramétriques y compris les arbres de décisions) adaptés à ce jeu de données.
2. On décide de retenir un modèle de régression logistique. Proposez 3 écritures rigoureuses de ce modèle.
3. Quelle méthode utilise-t-on pour estimer le modèle précédent ?
4. Que signifie l'algorithme IRLS ? A quoi sert-il ?
5. Lors de votre estimation, un collègue (non statisticien) vous propose d'utiliser la méthode des moindres carrés ordinaires pour estimer ce modèle, quelle sera votre réaction (arguments à l'appui) ?
6. Existe-t-il des valeurs manquantes, aberrantes et extrêmes dans les données ?
7. Réalisez une analyse descriptive des variables y compris la variable défaut (statistiques descriptives des variables, croisement entre les variables explicatives, croisement entre la variable à expliquer et les variables explicatives avec les indicateurs statistiques appropriés).
8. Quelles sont les conséquences de la multicolinéarité ?
9. Implémentez les modèles de régression logistique suivants :
   - $y = f(WCTA, RETA)$
   - $y = f\ (\ METL, EBIT_{TA})$
   - $y = f\ (WCTA, RETA, EBIT_{TA}, STA)$
   - $y = f(WCTA, RETA, EBIT_{TA}, METL, STA)$

   Interprétez pour chaque modèle les coefficients estimés. Donnez les courbes ROC associées à chaque modèle ainsi les aires sous ces courbes.

10. Quel modèle choisir selon les critères du BIC/SC et de l'AIC. Vous rappellerez la définition de ces critères et leur logique de construction
11. On décide de retenir le modèle avec 5 variables. Quelle est la probabilité de défaut d'une nouvelle entreprise ayant les caractéristiques suivantes : (WCTA = 0.6, RETA= 0.25, EBIT_TA = 0.45, METL= 0.05, STA= 0.72s)

**Partie B** : régression logistique sur variables discrétisées

Nous décidons d'implémenter le même modèle non plus sur les variables quantitatives directement mais dorénavant sur des variables discrétisées.

1. Quels sont les avantages d'une telle approche ?
2. Pourquoi sommes-nous obligés de poser des contraintes d'identifiabilité avant une telle estimation.
3. Proposez une discrétisation des 5 variables quantitatives avec les contraintes suivantes : 2 classes au moins et 3 classes au plus pour chaque variable.
4. Analysez la corrélation entre les différentes variables discrétisées à partir d'indicateurs adaptés.

**Projet de Scoring**
**M2 IMSD – Université Paris Saclay**
**Date limite : 15 Février**

5. Estimez un modèle de régression logistique sur les variables quantitatives discrétisées avec une méthode pas à pas. Interprétez les paramètres et les différents indicateurs de performance (AIC, BIC/SC, courbe ROC, indice de Gini,…).
6. Comparez les différents modèles estimés lors de la question précédente et sélectionnez le meilleur modèle selon le critère de votre choix.
7. Implémentez une grille de score sur une échelle de 1000