Financial Fraud Detection

Nom : **BANULESCU-RADU** Prénom : **Denisa** Année : **M2** Semestre : **9**

Nature : **CM** Volume horaire : **12 H** ECTS / Coef : **2**

Prérequis

- ✓ Notions d'économétrie linéaire et des variables qualitatives
- ✓ Bases en statistiques, probabilités et optimisation
- ✓ Connaissance des méthodes de machine learning (supervisé & non supervisé)

Résumé

Ce cours a pour objectif de former les participants aux méthodes économétriques et aux techniques d'apprentissage automatique, supervisées et non supervisées, appliquées à la détection de la fraude financière. Après une introduction aux principales typologies de fraude, l'accent sera mis sur l'analyse et le traitement des bases de données contenant des observations frauduleuses.

Deux grandes catégories de modèles seront étudiées, ainsi que leurs mesures de performance :

- **Modèles de prévention (non supervisés)** : conçus pour identifier, en amont, les transactions ou clients présentant des comportements atypiques.
- **Modèles de détection (supervisés)** : construits à partir de données historiques afin de classer les nouvelles observations comme frauduleuses ou non.

Un enjeu central est le caractère rare de la fraude : les bases de données utilisées sont très volumineuses mais fortement déséquilibrées (par exemple, les fraudes à la carte de crédit représentent souvent moins de 0,5 % des transactions). La dernière partie du cours portera ainsi sur les méthodes permettant de corriger ce déséquilibre et d'améliorer la robustesse des modèles.

Le cours sera complété par la réalisation d'un projet pratique mettant en œuvre les techniques étudiées sur une base de données spécifique.

Objectifs

- Comprendre la définition et les principales typologies de fraude financière
- Identifier et appliquer les techniques analytiques utilisées pour la détection de la fraude
- Analyser et traiter des bases de données contenant des cas de fraude
- Sélectionner les variables pertinentes pour la prévention et la détection de la fraude
- Mettre en œuvre des méthodes d'apprentissage automatique supervisées et non supervisées adaptées à la détection de la fraude
- Évaluer des modèles économétriques et de machine learning à l'aide de mesures de performance adaptées (classification et régression)
- Utiliser les principales méthodes de rééchantillonnage pour traiter les bases de données déséquilibrées (oversampling, undersampling, SMOTE, etc.)
- Réaliser une étude de cas pratique appliquée à la détection de la fraude

Bibliographie

- ✓ Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons.
- ✓ Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- ✓ Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). Learning from imbalanced data sets. Springer.
- ✓ He, H. and Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons.

PLAN

- **Chapter 1 - Introduction**

- **Chapter 2 - Data**
  - 2.1. Data typologies and sources
  - 2.2. Operations on data

- **Chapter 3 - Descriptive analytics for fraud detection – Unsupervised learning**
  - 3.1. Outlier detection
  - 3.2. Clustering

- **Chapter 4 - Predictive analytics for fraud detection – Supervised learning**
  - 4.1. Linear regression
  - 4.2. Logistic regression
  - 4.3. Decision trees
  - 4.4. Ensemble methods: bagging, boosting, random forest

- **Chapter 5 - Predictive models for skewed datasets**
  - 5.1. Undersampling
  - 5.2. Oversampling
  - 5.3. Adjusting posterior probabilities
  - 5.4. Cost-sensitive learning

- **Chapter 6 – Evaluation of predictive models for fraud detection**
  - 6.1. Data splitting
  - 6.2. Performance measures for classification models
  - 6.3. Illustration

- **Chapter 7 - Cost-sensitive learning for fraud detection**
  - 7.1. Cost matrix
  - 7.2. Cost-sensitive Logistic regression
  - 7.3. Cost-Sensitive evaluation metrics
  - 7.4. Illustration