

Détection de la Fraude

Nom : **BANULESCU-RADU**

Prénom : **Denisa**

Année : **M2**

Semestre : **10**

Nature : **CM**

Volume horaire : **12 H**

ECTS / Coef : **2**

Prérequis	<ul style="list-style-type: none">✓ Notions d'économétrie linéaire✓ Notions d'économétrie des variables qualitatives✓ Connaissance des méthodes de machine learning
Résumé	<p>Ce cours a pour objectif de former les participants aux méthodes économétriques et d'apprentissage automatique supervisée et non-supervisée dans le contexte spécifique de la détection de la fraude financière. Nous commençons par la définition et la présentation de différentes typologies de fraude, pour se concentrer par la suite sur l'analyse et le traitement spécifique des bases de données contenant des observations frauduleuses. Deux catégories de modèles de détection de la fraude, ainsi que leurs mesures de performance, seront ensuite présentées :</p> <ul style="list-style-type: none">(i) les modèles de prévention / non supervisés conçus pour détecter les fraudes avant qu'elles ne se produisent. Ces modèles identifient les transactions ou les clients les plus différents d'une norme donnée.(ii) les modèles de détection / supervisés conçus pour détecter les cas de fraude ex-post. Dans ce cas, des enregistrements antérieurs sont utilisés pour construire des modèles permettant d'affecter les nouvelles observations à l'une des deux classes. <p>La fraude est un évènement rare ce qui fait que les bases de données utilisées pour la détection de la fraude sont énormes, mais très déséquilibrées (e.g., les études de détection des fraudes à la carte de crédit indiquent un taux de fraude inférieur à 0,5%). Dans la dernière partie du cours, l'attention portera sur les principales corrections qui doivent être appliquées pour résoudre ce problème.</p> <p>Le cours sera complété par la réalisation d'un projet de mise en œuvre des techniques acquises sur une base de données spécifique.</p>
Objectifs	<ul style="list-style-type: none">• Définition et types de fraude financière• Présentation des techniques analytiques utilisées pour la détection de la fraude• Analyse de bases de données contenant des cas de fraude et leur traitement• Sélection des variables pertinentes dans le cadre de la prévention/détection de la fraude• Présentation des méthodes d'apprentissage automatique supervisées et non-supervisées utilisées dans la détection/prévention de la fraude• Évaluation des modèles économétriques et d'apprentissage automatique, en insistant sur les mesures de performance des modèles de classification et de régression• Présentation des principales méthodes utilisées pour le ré-échantillonnage des bases de données déséquilibrées (<i>imbalanced datasets</i>) : oversampling, undersampling, SMOTE, etc. <p>Réalisation d'une étude de cas pratique</p>
Bibliographie	<ul style="list-style-type: none">✓ 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. Support vector machine
 - 4.5. Ensemble methods: bagging, boosting, random forest
- **Chapter 5 - Evaluating predictive models**
 - 5.1. Performance measures for classification models
 - 5.2. Performance measures for regression models
- **Chapter 6 - Predictive models for skewed datasets**
 - 6.1. Oversampling
 - 6.2. Undersampling
 - 6.3. SMOTE