**PROJET MACHINE LEARNING**

# Tarification automobile à l'aide de modèles de machine learning

## Objectif

Une société d'assurance souhaite affiner sa capacité à individualiser les tarifs auto, tout en respectant au mieux le principe de mutualisation des risques (à la base de l'assurance). Son objectif est de faire payer à chaque assuré son « juste prix » afin de les fidéliser tout en maintenant l'équilibre technique.

Dans ce contexte, cette société d'assurance veut explorer de nouvelles approches de modélisation en recourant aux algorithmes de machine learning. L'objectif est de **construire un modèle qui prédit, pour chaque assuré, sa probabilité de déposera une réclamation au cours de la prochaine année**. Ainsi, une prédiction la plus précise possible permettra à l'assureur d'adapter davantage ses prix en l'adaptant au risque potentiel de chaque assuré.

## Données

Vous disposez de deux fichiers au format CSV :

- **base_train.csv** : données d'apprentissage (416 648 obs.) pour construire les modèles ; chaque ligne correspond à un assuré auto
- **base_test.csv** : données de test (178 564 obs.) permettant de mesurer la qualité prédictives des modèles

Chaque ligne de ces 2 fichiers correspond à un assuré auto. Toutes les variables de ces bases ont été anonymisées (pas de signification des variables). L'objectif est d'optimiser la performance prédictive des modèles.

Les bases TRAIN et TEST contiennent 59 variables dont :

- ID est l'identifiant de l'assuré
- TARGET représente la variable à prédire : la valeur 1 indique qu'une réclamation a été déposée par le client ; la valeur 0 indique qu'aucune réclamation n'a été déposée.
- 57 prédicteurs potentiels anonymisés. Dans le nom des variables, on distingue des familles de données indiquées par la présence des termes "**ind**", "**reg**", "**car**" et "**calc**". De plus, les noms des variables incluent le suffixe "**bin**" pour indiquer variables binaires, "**cat**" pour indiquer les variables catégorielles. Les variables sans ces désignations sont soit continues, soit ordinales. Enfin, les valeurs NA indiquent que l'information est manquante pour l'assuré.

## Démarche à suivre

Voici les opérations attendues dans le cadre du projet :

1. **Import et analyse des données.**
   - Importation des données sous R (ou Python)
   - Analyse exploratoire des données
   - Construction, transformation des données si besoin

2. **Modélisation du risque de réclamation (suite à un accident, un vol…)**
   - **Estimer la probabilité de risque de réclamation** (TARGET) à l'aide des variables disponibles et/ou transformées. Vous utiliserez différentes approches de machine learning en testant et confrontant des approches « simples » (k-NN, régression logistique*, arbre de décision) et des approches « avancées » (SVM, réseau de neurones, forêts aléatoire, gradient boosting). Vous commenterez votre démarche et vos choix.

     (*) Dans le cas de la régression logistique, vous devrez vous assurer que :
     - Chaque modalité des variables catégorielles doit contenir au moins 5% de la population d'apprentissage
     - Les corrélations entre les variables explicatives du modèle ne doivent pas dépasser un certain seuil (à définir par vos soins)
     - Chaque coefficient du modèle final doit être significatif au seuil de 10% et de signe cohérent par rapport au niveau de risque de réclamation.
   - **Mesurer et comparer la performance des modèles testés** sur les échantillons d'apprentissage (train) et de validation (test). Vous présenterez les indicateurs de performance retenus et commenterez les résultats.

Note d'attention :

- Selon les méthodes utilisées, vous devrez **être vigilant au risque de sur-apprentissage**
- Une difficulté du projet réside dans le **déséquilibre de la variable à prédire** (< 4%). A vous de voir s'il est pertinent ou pas de recourir à des méthodes de rééquilibrage.

## Modalité de l'épreuve

Ce projet s'inscrit dans le cadre des interventions de Valérie Monbet et moi-même.

Ce travail est à réaliser **par trinôme** (comme vous êtes 26 étudiants, il y aura un binôme) en utilisant le langage de votre choix : **R** ou **Python**.

Chacun des 9 groupes constitués tirera au hasard **deux algorithmes de machine learning à utiliser obligatoirement dans le cadre du projet**, avec la possibilité de tester d'autres modèles. Ce tirage au sort sera réalisé le 09 décembre et les travaux seront menés sur le temps des interventions de Valérie Monbet (10, 12, 17 & 18/12) ainsi que sur votre temps personnel.

Nous attendons une **présentation théorique des algorithmes à traiter obligatoirement** avec leurs modalités de mis en œuvre, leurs limites et avantages ainsi que les résultats obtenus sur le jeu de donnée fourni. Cela fera l'objet de la rédaction d'un document de type « article scientifique » et de la remise de vos codes R ou Python sous forme de markdown ou notebook.

Au-delà des restitutions écrites, nous prévoyons une **soutenance orale de vos travaux** devant l'ensemble des étudiants afin de partager la connaissance.

## Livrables

- Votre **article scientifique au format PDF (5 pages max.)** décrivant, à minima, les algorithmes de machine learning à traiter obligatoirement (tirage au sort) et les résultats obtenus sur le jeu de données. Vous indiquerez dans le nom du fichier transmis le nom de votre groupe.

- Votre **support au format PPT (8-10 diapos max)** qui sera présenté à **l'oral le 27 janvier 2020**. Vous indiquerez dans le nom du fichier transmis le nom de votre groupe

- Votre **code R ou Python sous forme de Markdown ou Notebook** : vos codes doivent être commentés, les plus automatisés et lisibles possible. Vous indiquerez dans le nom du fichier transmis le nom de votre groupe

Ce travail est à nous renvoyer **avant le jeudi 16 janvier 2020** aux adresses suivantes :

valerie.monbet@univ-rennes1.fr
dorothee.delaunay@mel.lincoln.fr

Pour toutes questions, vous pouvez nous joindre, Valérie Monbet et moi-même, par mail aux adresses indiquées ci-dessus.