Université de Lille **Méthodes d'Apprentissage**
M2 ISN - Data Sciences obouaziz@univ-lille.fr

TP. 4

Régression logistique - Données déséquilibrées

Un compte rendu est attendu pour ce TP :

- *seuls les fichiers aux formats .pdf ou .html seront acceptés (rendu sur Moodle)*
- *il n'est pas demandé d'inclure le code : c'est la présentation et l'interprétation des résultats qui sont importantes*
- *soyez concis (maximum 5 pages) : choisissez judicieusement les graphes à afficher, n'incluez pas de sorties brutes du logiciel*
- *à chaque figure et chaque tableau doit être associée une légende.*

Dans cet exercice on va travailler un jeu de données Telecom sur une campagne de ciblage marketing, on cherche à contacter les clients d'un opérateur téléphonique qui ont l'intention de se désabonner à un service. Pour essayer de cibler les individus ayant la plus forte probabilité de se désabonner (on a donc une variable binaire sur le fait de se désabonner ou non : Fichier Telecom_y.csv) on va mettre en place un algorithme de scoring des clients, en utilisant la base client existante (Telecom_x.csv pour les variables explicatives et Telecom_y.csv pour la variable à prédire) de l'opérateur dans laquelle les anciens clients qui se sont déjà désabonnés ont été conservés.

On va donc construire un modèle de régression logistique permettant d'expliquer et de prédire le désabonnement. Notre objectif est aussi d'extraire les caractéristiques les plus importantes de nos clients.

1. Charger le jeu de données, regarder quelques statistiques qui vous semblent pertinentes. Préparer un échantillon test et un échantillon d'apprentissage, justifier les tailles allouées à chaque échantillon.
2. Regarder l'aide la fonction `LogisticRegression` et l'appliquer au jeu de données d'apprentissage.
3. Calculer les odds-ratio et interpréter quelles variables ont un effet pour déterminer si un individu plus forte/faible probabilité de se désabonner ?
4. Regarder l'erreur associée sur la base de test. Que remarque-ton ? (On donnera la matrice de confusion associée).
5. Pour essayer de contourner l'effet des données déséquilibrées on va mettre en place trois stratégies :
   - Changer le seuil s qui associe un individu à la classe 1 si $\hat{\mathbb{P}}(Y = 1|X) \geq s$.
   - Sous-échantillonnage à l'aide de la fonction `NearMiss` : cette fonction supprime les observations de la classe majoritaire lorsque des observations

associées à des classes différentes sont proches l'une de l'autre. Inconvénient : le sous-échantillonnage de la classe majoritaire peut exclure des individus importants qui fournissent des informations nécessaires à la différenciation des deux classes

— Sur-échantillonnage à l'aide de la fonction `SMOTE` : crée de nouvelles instances de la classe minoritaire en utilisant les données. Les nouveaux individus créées ne sont pas une copie de ce qui existe mais proposent de nouvelles caractéristiques proches de ce qui a été observé dans la classe minoritaire.

Mettre en place ces 3 stratégies et comparer les performances obtenues. On expliquera le protocole mis en place pour cette comparaison et on discutera les avantages et les inconvénients de chaque méthodes.

6. Tracer les courbes ROC associées. Commenter.
7. Refaire les questions 1 à 4 sur le logiciel `R` avec la fonction `glm` (`family = binomial(logit)`). Comparer les sorties. D'un point de vue interprétation statistique quel logiciel est le plus facile à prendre en main ? Quelles sont les variables qui influencent la variable Y, comment ?