TP 2 - Master IREF : Prédiction du risque de crédit

Prérequis :

- Langage de programmation Python
- Utilisation des notebooks avec Jupyter
- Librairies recommandées : Pandas, Numpy, Scikit-learn, statsmodel

Objectif : Créer une grille de score basée sur la régression logistique et interpréter les résultats de la modélisation

Exercices :

1. **Encoder les variables candidates brutes en indicatrices imbriquées**

   Exemple :

   Soit A = {célibataire, marié, veuf}, un exemple de codage en indicatrices imbriquées :

	Indicatrices imbriquées	
A	***A*1**	**A2**
célibataire	0	0
marié	1	0
veuf	1	1

   Il est recommandé de trier l'ordre des modalités de la plus risqué à la moins risqué. Dans cet exemple, Célibataire est la modalité la plus risqué et Veuf la modalité la moins risqué.

2. **Modéliser une régression logistique sans variable (uniquement la constante)**

Statsmodel: from statsmodels.discrete.discrete_model import Logit (**recommandé**)
Scikit-learn : from sklearn.linear_model.LogisticRegression

Evaluer la performance en GINI du modèle. Que constatez-vous à propos du coefficient de l'intercept (constante) ?

3. Ajouter une première variable dans le modèle

- Sélectionner la variable avec l'IV la plus élevée
- Ajouter uniquement les modalités imbriquées

Etudier la table des coefficients obtenus par la régression logistique. Quels sont vos observations de ces premiers résultats ?

4. Implémenter une méthode par récurrence de sélection des variables

Afin de mieux contrôler l'ajout de variable dans le modèle et d'éviter des variables trop corrélées entre elles, nous allons utiliser le principe de l'IV pour sélectionner les prochaines variables à intégrer.

En vous appuyant sur le modèle précédemment entrainé, calculer le delta IV des variables candidates. Pour une variable candidate, il s'agit de calculer les effectifs réels et prédits à l'aide de la méthode predict_proba de votre modèle. Calculer l'IV réels et prédits et faite la différence par modalité. La somme des delta IV vous donnera l'importance de la variable à intégrer. Toutes variables < 0.05 peuvent être retirées du processus de sélection.

	BP	MP
Marié	100 / 90.5	20 / 29.5
Célibataire	50 / 48.3	25 / 26.7

5. Interpréter les résultats de la régression logistique

Une fois le processus d'entrainement terminé, vous pouvez analyser les résultats de la régression logistique.

- Performance du modèle (Gini, AUC)
- R-carré ajusté du modèle
- Critère d'information AIC / BIC
- Significativité des coefficients (Walt test)

6. Normaliser les coefficients de la régression logistique pour générer une grille de score

La grille de score se basera sur :

- Ratio de référence = 1
- Note de référence = 300
- Facteur multiplicatif = 2
- Nombre de points = 20

Afficher la grille de score finale, en pensant à transformer les variables imbriquées en modalité d'origine.

Observez la cohérence des coefficients et des taux de risques de votre grille de score, que constatez-vous ?