TP 1 - Master IREF : Prédiction du risque de crédit

Prérequis :

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

Données :

Vous travaillez pour une Banque vous demandant d'élaborer un outil automatisant le process d'octroi de crédit à ses futurs clients. La banque dispose d'un historique des crédits effectués par ses clients, avec des informations personnelles sur le client et s'il a fait défaut ou non pendant la période de remboursement du crédit. Les données contiennent les informations ci-dessous :

- GOOD_PAYER : 1 s'il n'y a pas eu de retard de remboursement, 0 sinon
- BAD_PAYER : 1 s'il n'y a pas eu de retard de remboursement, 0 sinon
- Mrev_Tit : Revenu du titulaire (en euros)
- Ressource : Ensemble des revenus du foyer (en euros)
- Charge : Charge total du foyer (en euros)
- Ancien_Prof_Tit : Ancienneté professionnel du titulaire (en année)
- RAV : Revenu après déduction de charge (en euros)
- Ratio_Ress_RAV : ration entre Ressource et RAV
- Tx_Edt : Taux d'endettement du crédit souscrit
- Age_Tit : Age du titulaire (en année)
- MCLFCHAB1 : Situation d'habitat du titulaire
- MCLFSITFAM : Situation familiale du titulaire
- CSP_TIT : Catégorie socio-professionnel du titulaire
- ZCOM_SR_MIMPOTS : Impôts mensuel du foyer (en euros)
- GEN_ACTIVE : Date d'octroi du crédit
- GEN_DEMAND : Date de demande du crédit

Objectifs : Analyser un jeu de données et préparer les données pour la modélisation

Exercices :

1. **Réaliser une analyse du jeu de données**
    a. Type des variables
    b. Nombre de valeurs manquantes
    c. Identification de valeurs aberrantes/erronées

Exemple : Situation Habitat → Variable catégorielle à 7 modalités / 191 valeurs manquantes

N'hésitez pas à utiliser des librairies graphiques (matplotlib, plotly, seaborn)

2. **Créer une table d'analyse marginale par variable**
    a. Nombre de client
    b. Nombre de client en défaut/ non-défaut
    c. % de client en défaut / non-défaut
    d. % de client
    e. Taux de risque

Exemple : Situation Habitat

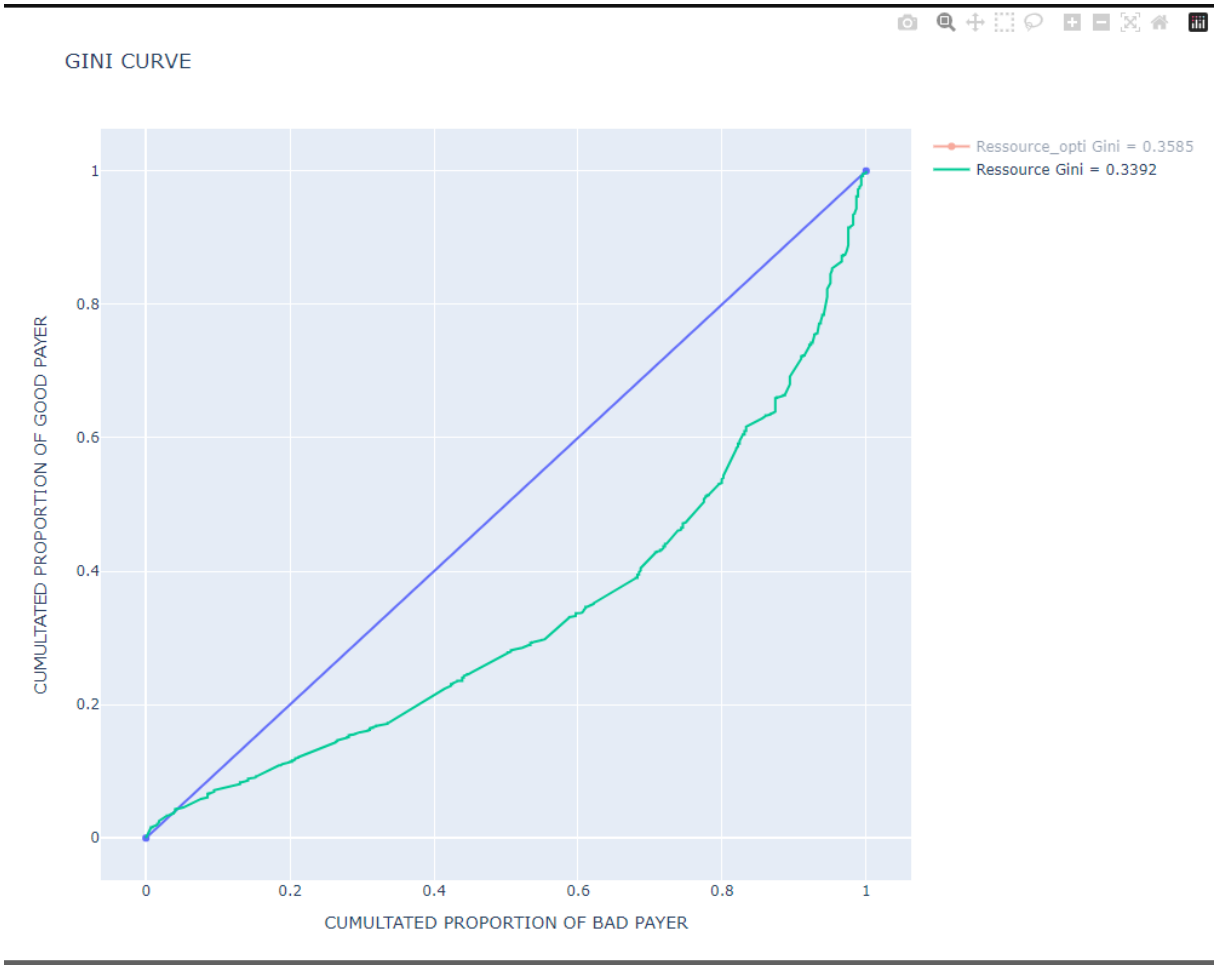
	Nbr Client	Nbr BP	Nbr BP	Taux de risque
Marié	100	80	20	20%
Célibataire	200	100	100	50%

3. **Implémenter une fonction calculant le poids d'évidence et l'information value d'une variable**

Vous pouvez vous référer à cette formule : Weight of Evidence (WOE) and Information Value (IV) Explained (listendata.com)

4. **Créer une fonction calculant le GINI d'une variable numérique et générer la courbe de concentration**

Utiliser la formule de Brown vu en cours ou bien appuyer vous sur la formule AUC (Area Under the Curve) → GINI = 2 * AUC - 1



5. **Analyse des variables**
    a. Générer un histogramme des taux de risque et des volumes de chaque modalité par variable
    b. Générer un graphique sur l'évolution des volumes et des taux de risques dans le temps par variable

    Commenter les différents graphiques générés et évaluer l'IV et le Gini des variables

6. **Créer une fonction générant plusieurs échantillons de données**

    Entraînement, validation et test (60%, 20%, 20%)

    Le package scikit-learn possède une méthode permettant d'automatiser ce processus (train_test_split), quelle différence statistique constater-vous entre les échantillons générés par le package et les vôtres ?

7. **Traitements des variables qualitatives**

    Veillez à respecter les contraintes de volumétrie d'un minimum de 5% de client au total par modalité et d'un minimum de 50 clients en défaut par modalité.

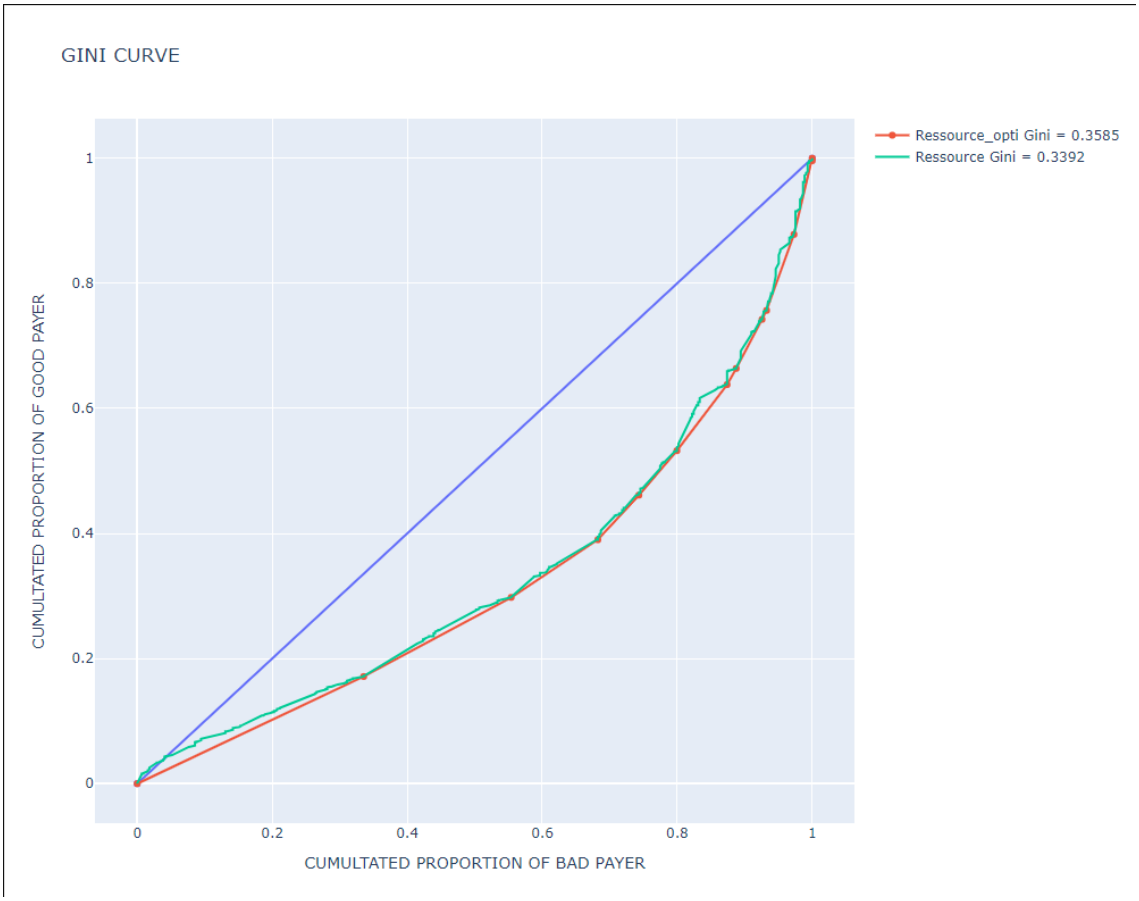
    Différentes possibilités s'offrent à vous :

- Regroupement par expertise/sens métier
- Regroupement par taux de risque ou poids d'évidence

8. Traitements des variables quantitatives

Créer une fonction d'optimisation des variables quantitatives se basant sur le GINI.

Dans l'exemple ci-dessous, il est possible de réduire le nombre de modalité de la variable et d'améliorer le GINI (la surface sous la courbe de la bissectrice) grâce aux calculs des pentes (dérivées) de cette courbe. La courbe verte représente la variable brute Ressource et la courbe rouge la variable Ressource optimisée par le GINI.



9. Identification des variables candidates pour un modèle de Scoring

a. Etablir un classement des variables par GINI et par IV
b. Créer une matrice de corrélation basée sur le V de Cramer (Chi2)
c. Sélectionner les variables avec un IV > 0.05 et une corrélation < 0.8

Vous avez maintenant à disposition un jeu de donnée prêt à être entraîné. Le TP 2 se focalisera sur l'élaboration d'une grille de score à partir de la régression logistique.