Fortuneo

Etude de cas

Construction d'un modèle de machine Learning permettant de prédire si un client sera intéressé ou pas par un nouveau produit d'assurance automobile.

N'DRI Jean Philippe Cedric

+33 6 12 77 53 36
ndricedric0@gmail.com

CONTEXTE

Dans le cadre du lancement d'un nouveau produit d'assurance automobile, une compagnie d'assurance santé souhaite mettre en place un modèle de machine Learning afin de prédire si les clients seront intéressés par ce produit.
Le but de notre étude est de construire ce modèle d'appétence pour permettre à l'entreprise de planifier en conséquence sa stratégie de communication pour les clients et d'optimiser son modèle commercial.

ANALYSE DE NOTRE BASE DE DONNÉES

Une première analyse de notre base de données nous a permis de remarquer que notre base d'apprentissage (data_train) était fortement déséquilibrée.
Pour résoudre le problème de déséquilibre dans notre base de données, nous avons envisagé les solutions suivantes :

Le Sous-échantillonnage de la classe majoritaire (Undersampling)
Cette méthode consiste à réduire le nombre d'observations de la (des) classe(s) majoritaire(s) afin d'arriver à un ratio classe minoritaire/ classe majoritaire satisfaisant.

Le Sur-échantillonnage de la classe minoritaire (Oversampling)
Elle consiste à augmenter le nombre d'observations de la (des) classe(s) minoritaire(s) afin d'arriver à un ratio classe minoritaire/ classe majoritaire satisfaisant.

SMOTE
Elle fonctionne en créant des échantillons synthétiques à partir de la classe minoritaire au lieu de créer de simples copies.

SMOTEENN
Combine à la fois le sur-échantillonnage SMOTE de la classe minoritaire
et le sous-échantillonnage la classe majoritaire grâce à la méthode des plus proches voisins.

Utilisation d'un modèle pénalisé (Cost-Sensitive Algorithms)

Cette méthode consiste à imposer un coût supplémentaire au modèle pour les erreurs de classification commises sur la classe minoritaire pendant la formation. Ces pénalités peuvent biaiser le modèle pour qu'il accorde plus d'attention à la classe minoritaire

Pour cette étude, la méthode **SMOTEENN** sera c'elle privilégiée.

LES DIFFERENTES ETAPES VERS LA MISE EN PLACE DU MODÈLE

- Chargement de la base de donnée et recodage des variables
- *Le recodage des variables a été réalisé comme suit :*
    - *Gender: Male: 1,Female: 0*
    - *Vehicle_Damage: Yes: 1,No: 0*
    - *Vehicle_Age: < 1 Year: 1, 1-2 Year: 2, > 2 Years: 3*

- Rééchantillonnage de la base de données grâce à la méthode SNOTEENN. La fonction ci-dessous est celle utilisée pour réaliser le rééchantillonnage.

```
def Transformation(X,Y):
  #define sampling strategy
  sample = SMOTEENN(sampling_strategy=0.5)
  #fit and apply the transform
  X_over, y_over = sample.fit_resample(X, Y)
```

- Evaluation de plusieurs classificateurs
- Sélection des meilleurs classificateurs
- Entrainement du modèle final
- Création de pipeline
- Evaluation avec les données de test

LES RESULTATS

Avec une précision de 89.35% et un f1-score de 92.92%, le modèle qui sera utilisé par la suite est le KNN.
Le modèle arrive à prédire que 78% des clients (99502) ne sont pas favorables au nouveau produit tandis que 22% sont favorables.