APPLICATIONS DE L'ANALYTICS EN GESTION DES RISQUES

Anani OLYMPIO, CNP Assurances, Laboratoire SAF, Université Lyon 1

INTRODUCTION

Les thèmes de ma recherche portent sur les méthodes de détection et d'analyse des risques (comportements et biométriques). Par exemple, nous étudions:

1. **Risque comportemental :** modélisation des profils d'attitudes face au risque et face au modèle à partir des algorithmes de Logique Floue (**Fuzzy C-Means**)
2. **Risque biométrique** : modélisation des durées de sinistres et des montants cumulé sdes sinistres à payer pour le provisionnement non-vie par Machine Learning (**Random Forest Censored** et **Gradient Tree Censored Boosting**)

PROBLEMATIQUE

Dans le cadre du provisionnement non-vie tête par tête, nous proposons deux procédures permettant d'estimer la distribution conditionnelle d'une variable censurée comme alternatives à la méthode « *Tree based censored regression* » de Olivier Lopez et al (2016). Cette dernière est une modification de la méthode CART (Classification And Regression Trees) de Léo Breiman et al. (1984) dont l'une des limites est l'instabilité.

Comment assurer la performance et la stabilité des résultats ?

BIBLIOGRAPHIE

- Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984: Classification and regression trees. Wadsworth Books, 358.
- Breiman, L., 1996: Bagging predictors. Machine learning, 24 (2), 123--140.
- Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32. doi :10.1023/A: 1010933404324
- Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. Computational and Mathematical Methods in Medicine, 2013, 873595. http://doi.org/10.1155/2013/873595
- O. Lopez, X. Milhaud, P. Thérond Tree based censored regression with applications to insurance, Electronic Journal of Statistics Vol. 10 (2016) p. 2685-2716

NEXT STEP

- Comment modéliser les valeurs atypiques ?
- Amélioration de l'interprètabilité des résultats
- Application au risque de dépendance

MODELES

Random Forest

▷ **A RETENIR**

Avantage : Réduction de la volatilité du modèle, pris en compte de plus de features.

- Introduit par Breiman (2001) : variante du *bagging* utilisant CART.
- *Bagging* :
  - ⇒ Bootstrap : On tire aléatoirement un échantillon des observations.
  - ⇒ On crée un modèle sur cet échantillon puis on répète M fois.
  - ⇒ *Aggregating* : On fait une moyenne des modèles obtenus.
- Le *Random Forest* est un *bagging* amélioré utilisant CART.

Formellement Soit $(\psi_k)_{k=1\cdots N}$ une suite d'arbre de décision de variance σ et de corrélation ρ alors la variance de l'aggrégation de ces modèles qu'on notera ψ est :

$$\mathrm{Var}(\psi) = \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{N}$$

Gradient Tree Boosting

▷ **A RETENIR**

- ⇒ Le modèle original a été proposé par *R.Schapire* et *Y.Freund*.
- ⇒ **Gradient Boosting = Boosting + Descente de Gradient** .
- ⇒ Le *Boosting* : Addition de plusieurs modèles de régressions faibles.
- ⇒ Le Gradient Tree Boosting utilise CART pour créer les modèles de régressions faibles.

▷ **Formellement** l'idée est la suivante soit $X = (x_1, \cdots, x_n)$ et $Y = (y_1, \cdots, y_n)$:

$$H(X) = Y - F_1(X)$$

Dans le cas d'une régression, on pose comme fonction de coût :

$$L(y, F(x)) = \frac{1}{2}(y - F(x))^2$$

$$\frac{\partial(L(Y, F(X)))}{\partial F(X)} = F(X) - Y = H(X)$$

Exemple

On suppose qu'on a une variable *target* $Y \in R^+$ à prédire et $X = (X_1, X_2, X_3)$ variable explicative. On suppose aussi que l'on a les poids de Kaplan-Meier de la durée Y qu'on notera W_i.

Survival function

On appelle fonction de survie $f : t \to \mathbb{P}(T \geq t)$

Censure

Soit T et C deux variables aléatoires positives. On appelle une variable de censure (à droite) la donnée de deux éléments :

- $Y = \min(T, C)$
- $\delta = 1_{T \leq C}$

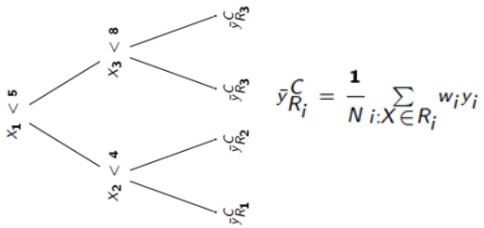
- Les estimateurs classiques ne marchent pas en cas de censure.
- Pour estimer une fonction de survie dans le cas de données censurées à droite :

$$\hat{S}(t) = 1 - \sum W_{i,n} 1_{Y_i < t}$$

Avec W_i les poids de Kaplan-Meier définis par :

$$W_{(i),n} = \frac{\delta_i}{n} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_i}$$

l'estimation de la moyenne n'est plus équipondérée mais on donne plus d'importance aux durées longues et observables.



$$\bar{y}_{R_i}^{C} = \frac{1}{N} \sum_{i:X \in R_i} w_i y_i$$

APPLICATIONS ET RESULTATS

- Calculs des durées et montants **sinistre non-clos**
- Portefeuille de prévoyance collective :

Caractéristiques :

- ✓ Risquess : arrêts de travail et maladie
- ✓ âge moyen survenance: 46 ans,
- ✓ durée moyenne sinistre : 3,5 mois
- ✓ Prestations moyenne : 3K€

Principaux résultats :

- ✓ Résultats Random Forest Censored et Gradient Tree Censored Boosting sont équivalents
- ✓ Globalement, une bonne qualité des prédictions avec l'approche Machine Learning
- ✓ Légères sous-estimations des prédictions sur certaines tranches d'âges liées à une faible volumétrie de données

Figure – Durée moyenne en fonction du risque×Age

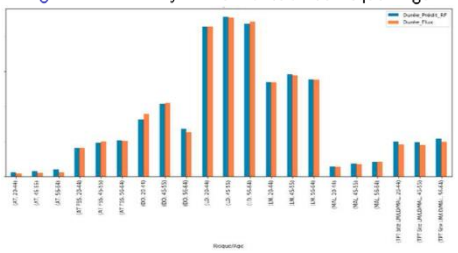


Figure – Prediction vs Réel sinistre non clos : Somme