UTILISATION DU MACHINE LEARNING POUR L'ANALYSE DE SURVIE

MOTS CLES : APPRENTISSAGE SUPERVISE - ANALYSE DE SURVIE - TIME-TO-EVENT MACHINE LEARNING - DATA SCIENCE - MODELE DE COX

Cet article a été rédigé par le Data Lab du cabinet Galea. Il a été enrichi grâce aux réflexions de Stéphanie BRUGIRARD, Responsable de l'équipe Data & Risks Monitoring de la direction de l'Actuariat Pôle Vie de Crédit Agricole Assurances qui permet la mise en perspective pratique de la théorie.

CLASSIFICATION, RÉGRESSION ET *RANKING*

Les tâches d'apprentissage supervisé sont habituellement utilisées pour répondre à deux types de problème : la classification et la régression. L'objet de cet article est de présenter une troisième famille de problèmes pour lesquels l'apprentissage supervisé s'avère extrêmement utile en sciences actuarielles et notamment en analyse de survie. On parlera de problèmes d'ordonnancement ou de *ranking*, permettant de traiter des tâches qui font intervenir une relation d'ordre. Une première formalisation du traitement d'un problème d'analyse de survie en présence de données censurées comme un problème de *ranking* est publiée dans un article de 2008[i] .

La plupart des algorithmes d'apprentissage statistique familiers des *data scientists* sont également adaptables à ce genre de problématiques et peuvent par exemple compléter les modèles de durée mis en œuvre de manière classique dans plusieurs domaines : étude du risque de mortalité et de longévité, étude des durées d'arrêt de travail, de chômage, construction de loi de rachats, ou encore étude de la valeur client.

Les consultants de Galea sont à votre disposition pour plus de précisions sur le contenu de cette note et pour vous accompagner dans la mise en place de modèles de *machine learning* adaptés à l'analyse de survie. Et plus généralement des modèles nécessitant la double compétence actuaire/*data scientist*.

L'ANALYSE DE SURVIE

NOTATIONS

L'analyse de survie est une branche des statistiques qui a pour objectif de modéliser des durées avant l'occurrence d'un évènement d'intérêt. Soit T la variable aléatoire réelle positive qui représente cette durée. Afin de rendre l'énoncé plus explicite, l'évènement d'intérêt sera par exemple le décès d'un assuré pour un contrat d'épargne retraite versant une rente jusqu'à la date du décès. En analyse de survie, comme pour les études de durée, une représentation de la variable d'intérêt par la fonction de survie S est préférée à une représentation par la fonction de répartition F.

$$S(t) = 1 - F(t) = P(t < T)$$

Il s'agit d'une fonction décroissante, telle que

$$S(0) = 1 \text{ et } \lim_{t \to +\infty} S(t) = 0.$$

La fonction de risque instantanée, ou fonction de hasard est une autre variable d'intérêt. Elle est définie comme la variation instantanée du logarithme de la fonction de survie :

$$h(t) = -\frac{d}{dt}\ln\big(S(t)\big)$$

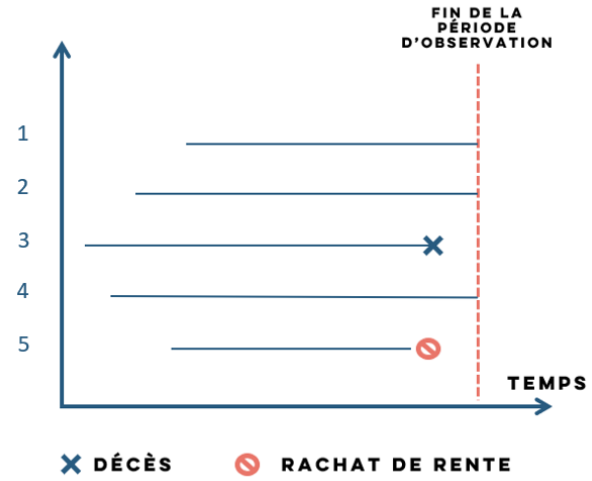
LA CENSURE DES DONNÉES

Une des spécificités en analyse de survie est la gestion de données « incomplètes ». On parle de censure lorsque l'évènement d'intérêt n'est pas observable sur la période d'observation. Dans cet exemple, la donnée sera censurée si :

// L'assuré décide de racheter son contrat au cours de la période d'observation ;

// L'assuré est encore vivant à la fin de la période d'observation.

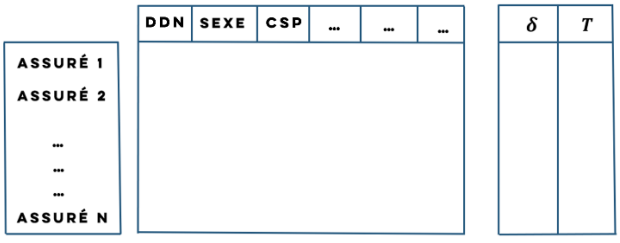
Ainsi, à la fin de la période d'observation, il est impossible de savoir si un individu ayant racheté son contrat avant la fin de la période d'observation est encore en vie ou non. Cette information permettant de faire la différence entre les assurés décédés, pour lesquels le risque est survenu, et les autres est primordiale afin d'étudier le risque de longévité.



A titre d'exemple, dans la représentation ci-dessus, les lignes de données 1, 2, 4 et 5 sont censurées car nous n'avons pas connaissance de la survenance d'un décès au cours de la période d'observation.

LES DONNÉES D'APPRENTISSAGE

Opérationnellement la présence du type de données spécifiques à l'analyse de survie se traduit par une structure de base de données d'entrainement bien particulière :



Cette base de données, contient des variables explicatives sur l'assuré, telles que la date de naissance (permettant de déduire l'âge et la génération), le sexe, la catégorie socio professionnelle, *etc.* La particularité réside dans la variable de sortie qui est en fait un couple de variable :

// T représente une durée jusqu'à la fin de l'observation de l'assuré (survenance de l'évènement d'intérêt ou censure). Dans le cadre d'études de longévité ou de mortalité, comme dans notre exemple, T représente généralement la durée de vie, *i.e.* l'âge à la fin de l'observation.

// δ est une variable booléenne qui vaut VRAI (1) si l'assuré est décédé pendant sa présence dans le portefeuille, FAUX (0) sinon.

Les algorithmes d'apprentissage supervisé classiques ne permettent pas de traiter une telle variable de sortie. Il est ainsi nécessaire de modifier ces algorithmes afin de les adapter aux problèmes d'analyse de survie et aux données censurées.

Et comme pour tout problème d'apprentissage sur des données, il est important de garder en tête l'adage « *garbage in, garbage out* ». Cette étape clé de maitrise des données, nécessaire à toute étude réussie mais qui n'est pas encore un total acquis chez tous les acteurs de la place, a déjà été franchie au sein de Crédit Agricole Assurances : « *Les données collectées par Crédit Agricole Assurances sont volumineuses et de qualité. Les infrastructures permettant leur stockage et traitement sont opérationnelles.* »

APPRENTISSAGE ET ANALYSE DE SURVIE

LE MODÈLE DE COX POUR LA FONCTION h

Avec le modèle de Makeham, et les estimateurs de Kaplan-Meier et Nelson-Aalen, pour ne citer qu'eux, le modèle de Cox fait partie de ces outils bien connus par la communauté actuarielle. C'est un modèle semi-paramétrique également très utilisé en biostatistique, pour lequel l'estimation des paramètres peut se faire par apprentissage. Il est utilisé dans le cas de portefeuilles hétérogènes pour tenir compte de co-variables et différencier la mortalité par sous-populations, par exemple en assurance emprunteur. Il

suppose que la fonction de risque instantanée h prend la forme :

$$h(t|x_{i,1}, x_{i,2}, \dots, x_{i,p}) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{i,j}\beta_j\right)$$

où $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ est le vecteur des p variables explicatives pour l'assuré i. h_0 est la fonction de hasard instantanée de référence (ou *baseline*) supposée indépendante des variables explicatives. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ est le vecteur des paramètres du modèle, il donne un poids à chacune des variables explicatives. Une cordonnée j telle que β_j est de valeur absolue élevée indiquera que la j-ème variable explicative a un pouvoir explicatif fort sur la mortalité.

Pour estimer $\boldsymbol{\beta}$, on compare les assurés deux à deux pour faire disparaitre la dépendance en h_0.

$$\ln\left(\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)}\right) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\intercal}\boldsymbol{\beta}$$

C'est par cette comparaison d'individus deux à deux que se fait le lien entre les modèles d'apprentissage en analyse de survie et les modèles d'apprentissage d'ordonnancement.

CONSTRUCTION D'UNE RELATION D'ORDRE ET MODIFICATION DU PROBLÈME D'OPTIMISATION

Mathématiquement, comparer des objets deux à deux consiste à construire une relation d'ordre sur l'espace de ces objets. L'objectif de la comparaison entre deux individus est dans notre cas de déterminer lequel des deux vivra le plus longtemps. La relation d'ordre pour la tâche d'analyse de survie est construite ainsi :

$$\boldsymbol{x}_i \preccurlyeq \boldsymbol{x}_j \iff (T_i < T_j \; et \; \delta_i = 1)$$

Pour affirmer que l'assuré j vivra plus longtemps que l'assuré i, il faut que l'assuré i soit décédé et que l'assuré j ait déjà vécu plus longtemps que l'assuré i. C'est le seul cas de comparaison possible. En effet si les deux assurés sont encore vivants, la comparaison n'est plus possible. Si $T_i \geq T_j$ et $\delta_i = 1$, l'assuré j peut décéder avant que T_j ne dépasse T_i, la comparaison n'est alors pas possible. Finalement la tâche d'apprentissage consiste en un problème d'optimisation de la log-vraisemblance, cela revient à chercher :

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} \delta_i \left[\boldsymbol{x}_i^{\intercal}\beta - \log \sum_{j \in \mathcal{R}_i} \exp(\boldsymbol{x}_j^{\intercal}\beta)\right]$$

La fonction h_0 est ensuite calculée comme un ajustement entre les données réelles et les données estimées avec le paramètre $\hat{\beta}$.

La fonction de risque instantané dépendant des variables explicatives, il est ainsi possible de construire des courbes de survie et de hasard cumulé individualisées.

Une fois le modèle entrainé il est possible d'utiliser les prédictions du modèle pour calculer des espérances de vie ou encore construire des tables de mortalité. Les résultats peuvent être exploités comme ceux produits par les modèles de durée plus traditionnels.

L'apport des techniques d'apprentissage statistique permet d'améliorer l'optimisation du problème pour des données en grande dimension mais également de généraliser le modèle de Cox par des modèles plus complexes permettant la prise en compte de non-linéarités. Il est ainsi possible d'utiliser des modèles à base d'arbre (*CART, Random Forest, XGBoost*) basés sur un critère de séparation adapté au *ranking* ou encore de modifier le terme linéaire de la fonction de risque instantanée de Cox par une fonction non linéaire correspondant à une architecture donnée de réseau de neurones.

PROGRAMMATION EFFECTIVE DES ALGORITHMES D'APPRENTISSAGE DE SURVIE

IMPLÉMENTATION SOUS R ET PYTHON

Il existe des implémentations des algorithmes classiques adaptées à l'analyse de survie. On peut par exemple citer la bibliothèque Python scikit-survival[ii] qui propose des implémentations

de *Random Survival Forest*, *Gradient Boosted Models* et *Support Vector Machine*.

Sur R, de nombreuses librairies fréquemment utilisées sont également adaptées à l'analyse de survie, c'est le cas de **ranger**[iii] ou **randomForestSRC**[iv] pour l'implémentation de forêts aléatoires.

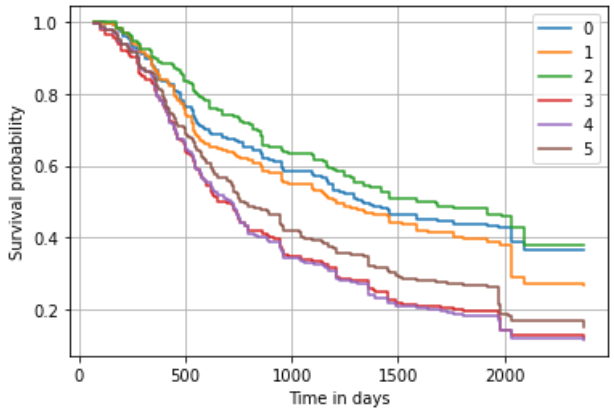
Il est également possible de faire appel à des réseaux de neurones afin de faire du *Deep Survival Analysis*. **Deepsurv**[v] (Python) et **Cox-Time**[vi] (R) peuvent alors être utilisées.

Lors de l'implémentation de ces modèles, il convient de respecter la méthodologie classique applicable en *machine learning*, ainsi que le processus de validation composé des principales étapes suivantes :

- Construction des bases d'apprentissage et de test ;
- Optimisation des hyperparamètres par validation croisée ;
- Evaluation du modèle.

ELÉMENTS D'ANALYSE

Les sorties de ce type de modèles sont assez homogènes parmi les librairies. Les résultats correspondent aux fonctions de survie et de hasard cumulé, construites par individu pour les points d'entrainement *ie* les durées non censurées observées (*cf.* graphique ci-contre, issu de la documentation Scikit-Survival).



La plupart des librairies proposent également comme prédiction, le « score de risque » calculé comme étant la somme, sur les points d'entrainement, de la fonction de hasard cumulé estimée. Cette valeur peut être difficile à interpréter mais permet le classement des individus par durée de vie telle qu'estimée par le modèle. L'espérance de vie, prédiction plus naturelle, peut également être obtenue en considérant l'aire sous la courbe de survie.

Concernant les métriques à utiliser pour l'évaluation et la comparaison de modèles, la plus classique est l'indice de concordance de Harrell (Harrell's C-index)[vii]. Parmi l'ensemble des paires comparables, on considère qu'une paire est « concordante » si la relation d'ordre entre les durées observées est conservée par les prédictions du modèle. L'indice de concordance correspond alors à la proportion de paires concordantes sur le nombre de paires permises. Cet indice facilement interprétable présente de nombreux avantages, il est adapté aux données censurées et son analogie avec l'*AUC* (*Area Under ROC Curve*) utilisée en classification le rend intuitif. Notons toutefois qu'il présente quelques défauts, il surestime la performance des modèles pour les taux de censures élevé et n'est pas approprié à l'étude de la pertinence des prédictions à un horizon fixé. D'autres métriques peuvent alors être utilisées en complément (*Brier Score, Time dependant AUC, …*)[viii].

CONCLUSION

Les algorithmes d'apprentissage statistique pour l'analyse de survie sont bien connus et maitrisés depuis plusieurs années, particulièrement dans le domaine médical, mais restent encore peu exploités en actuariat.

Au sein de Crédit Agricole Assurances : « *La priorité a été donnée aux applications de machine learning permettant la détection d'anomalies ou l'identification de variables pouvant prédire le comportement des clients. A ce jour, les durées restent modélisées grâce aux modèles traditionnels, leur modélisation par apprentissage sont à l'étude.* »

Des travaux menés à partir de ce type de modèles ont fait leurs preuves lors d'applications à la prédiction des départs à la retraite au sein du Data Lab Galea. Les résultats obtenus permettent de challenger les méthodes actuelles et de développer des compétences supplémentaires au sein du Data Lab Galea. Les performances de ces modèles apparaissent supérieures à celles obtenues à la fois avec les méthodes actuarielles classiques, et avec les approches de *machine learning* générales, non adaptées aux problèmes d'ordonnancement et à la censure des données. Les modèles d'apprentissage pour analyse de survie permettent une analyse à la maille individuelle. Les modèles actuariels classiques restent indispensables à la validation de ces modèles au global, dans un but d'explicabilité. Ainsi, ces nouveaux outils apportent de l'information de manière complémentaire aux méthodes classiques sans toutefois s'y substituer.

//galea

https://www.galea-associes.eu/

[i] Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., & Raykar, V. C. (2008). *On ranking in survival analysis: Bounds on the concordance index*. In *Advances in neural information processing systems* (pp. 1209-1216)

[ii] https://scikit-survival.readthedocs.io/en/stable/index.html

[iii] https://cran.r-project.org/web/packages/ranger/index.html

[iv] https://cran.r-project.org/web/packages/randomForestSRC/index.html

[v] https://pypi.org/project/deepsurv/

[vi] https://raphaels1.github.io/survivalmodels/reference/coxtime.html

[vii] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. *Evaluating the Yield of Medical Tests*. *JAMA.* 1982;247(18):2543–2546. doi:10.1001/jama.1982.03320430047030

[viii] https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html