# Modèles de prédiction interprétables pour processus non-stationnaires

CIFRE SAP-LIPN (UMR CNRS 7030)

Mme Hanane Azzag & M. Mustapha Lebbah

## 1 Contexte : la société SAP et les challenges

SAP est l'un des plus grands éditeurs de logiciels de gestion de processus métier au monde, et offre des solutions qui permettent un traitement des données et des flux d'informations efficaces au sein des organisations. Fondée en 1972, elle a d'abord été baptisée Systemanalyse Programmentwicklung (développement de programmes d'analyse de système), pour devenir SAP par la suite. De petite startup de cinq personnes, elle est passée à une entreprise multinationale de plus de 100 000 employés et plus de 440 000 clients dans 180 pays. Son siège social est basé à Walldorf en Allemagne.

Avec ses premiers logiciels, SAP R/2 et SAP R/3, SAP a établi la norme en matière de logiciels de planification des ressources de l'entreprise (ERP). Avec SAP S/4HANA, l'ERP passe au niveau supérieur en utilisant la puissance de la technologie In-memory pour traiter de grandes quantités de données et prendre en charge des technologies avancées telles que l'intelligence artificielle (IA) et le Machine Learning.

Les applications intégrées de l'éditeur connectent toutes les composantes d'une entreprise en une suite intelligente sur une plateforme numérique. Aujourd'hui, SAP compte plus de 215 millions d'utilisateurs Cloud, plus de 100 solutions qui couvrent toutes les fonctions métier et le portefeuille d'offres Cloud le plus fourni.

## 2 Les objectifs du sujet de recherche

Parmi les problèmes d'apprentissage traités par SAP, on distingue traditionnellement les tâches liées à l'apprentissage supervisé comme la classification/régression, bien adaptées à des processus stationnaires, et l'analyse et la prévision de séries temporelles, pour les processus non stationnaires. En classification/régression, la plupart des algorithmes classiques ne permettent pas d'extrapoler précisément la réponse à une variable au-delà du domaine rencontré en apprentissage. Le temps et les variables évoluant dans le temps sont donc généralement exclues du champ de la modélisation : une hypothèse classique est que le processus modélisé est suffisamment stationnaire pour que les données d'apprentissage soient représentatives du comportement à l'horizon temporel des prédictions souhaitées. En contrepartie, les algorithmes de classification/régression permettent une modélisation très fine des contributions de centaines de variables prédictives, incluant la prise en compte d'interactions complexes [LL17].

L'analyse des séries temporelles apparaît dans presque tous les domaines dont les variables dépendent fortement du facteur temps : anticipation d'utilisation de ressources, prévisions de ventes, de dépenses, ou d'abonnements, prévision de fréquentation de parcs d'attraction. La modélisation

de l'évolution d'un signal en fonction du temps est au cœur de l'analyse de séries temporelles, avec deux difficultés principales : détecter les ruptures dans les tendances et identifier des prédicteurs parmi les variables dont il est possible d'anticiper les futures valeurs (à titre d'exemple des événements récurrents comme les vacances scolaires). Les modèles de séries temporelles font en pratique intervenir peu de variables, avec des modèles additifs ou multiplicatifs simples, ignorant les interactions entre les variables.

Le sujet de recherche s'articule autour de deux aspects : (1) la construction d'un modèle robuste et sophistiqué pour le traitement de séries temporelles en présence de ruptures et (2) l'extension des modèles de classification/régression à l'extrapolation de tendances. Notre ambition est de briser la séparation traditionnelle entre classification/régression d'une part, et prévision de séries temporelles d'autre part, en construisant un modèle prédictif unifié intégrant le temps, les facteurs évoluant lentement dans le temps, ainsi que toutes les variables connues au moment d'une prédiction. Ce modèle doit :

— être assez complexe pour pouvoir apprendre les processus sous-jacents aux données.
— ne pas nécessiter une puissance de calcul exigeante.
— avoir des performances acceptables avec peu de données.
— être interprétable en un temps raisonnable.

Les arbres de décision et en particulier les Gradient Boosted Trees [Fri00 ; Mas+00 ; CG16] satisfont trois des quatre propriétés désirées. Cette famille d'algorithme est très souvent utilisée en 2020 pour les tâches de classification/régression, en particulier par SAP. L'entraînement des modèles est rapide par rapport aux réseaux de neurones (architectures profondes) et nécessite beaucoup moins de données pour avoir de bonnes performances. Enfin on peut extraire d'une forêt décisionnelle des interprétations locales, en temps polynomial [LL17 ; Lun+20]. Malheureusement, les arbres de décision sont performants en interpolation mais pas en extrapolation. En effet, la qualité des prédictions se dégrade dès que les valeurs prises par des variables sortent de la plage des valeurs apprises en entraînement, ce qui est souvent le cas des données qui dépendent du temps. Il s'agit là du principal obstacle à l'intégration du temps parmi les variables du modèle, dès lors qu'il s'agit de faire des prédictions à un horizon temporel situé dans le futur.

Nous proposons de remédier à ce problème en agissant sur les contributions au sens de Shapley [LL17]. Les valeurs de Shapley [Sha88] issues de la théorie des jeux permettent d'établir dans le cas d'un jeu coopératif une répartition équitable des gains de chacun des joueurs. Ce concept peut être étendu aux problèmes d'apprentissage statistique : si $\hat{f}$ est un modèle prédictif qui prend en entrée des données à plusieurs variables, on peut obtenir pour chaque donnée la contribution de chacune de ses variables à la prédiction finale. Par exemple, si le modèle permet de prédire le prix de l'immobilier en prenant en paramètre la surface en $m^2$, le nombre de pièces et la ville alors on peut obtenir, grâce aux valeurs de Shapley, la contribution de chacune de ces variables à la prédiction du prix. Si l'appartement a une surface de 50 $m^2$, 2 pièces, est localisé à Paris et qu'il coûte 500,000 euros, on peut savoir de combien chacune des variables a contribué, en euros, au prix final moins la moyenne des prix prédits. Dans le cas où la prédiction moyenne est de 300,000 euros, la surface de 50 $m^2$ peut avoir contribuer à hauteur de 100,000 euros, la localisation 50,000 euros et le nombre de pièces 50,000 euros, ce qui donne après addition le prix final. En effet, dans le cadre des valeurs de Shapley, la somme des contributions donne la prédiction finale moins la prédiction moyenne :

$$\hat{f}(x) - \frac{1}{N}\sum_{i=1}^{N}\hat{f}(x^i) = \sum_{i=1}^{n}\phi_i(x)$$

où $x \in \mathbb{R}^n$, $\phi_i(x)$ la contribution de $x_i$ et $(x^1, ..., x^N)$ les données d'apprentissage.

Les travaux de Lundberg et al. 2018 [LL17 ; Lun+20] ont permis de démocratiser l'utilisation des valeurs de Shapley sur des modèles tels les réseaux de neurones et les arbres de décision. Ceci rend possible l'utilisation de l'interprétabilité au sens Shapley des arbres de décision pour des tâches de prédiction avec des séries temporelles : un travail préliminaire effectué au cours d'un stage (Janvier-juin 2020) a montré que les valeurs non rencontrées en entraînement ont des contributions constantes et donc anormales. Il est alors envisageable d'extrapoler les contributions futures de certaines variables en se basant sur le passé ou en faisant des hypothèses sur leur devenir.

A notre connaissance, les méthodes d'attribution de gains n'ont pas été suffisamment exploré à la sélection de variables et l'apprentissage non supervisé. Il est possible d'estimer la contribution de chacune des valeurs des variables à l'erreur du modèle et on pourrait alors envisager d'enlever toutes les variables qui font augmenter l'erreur. Toutefois, l'erreur du modèle n'est en général pas homogène sur l'espace des variables, et la pertinence d'une variable peut ainsi dépendre du contexte. Une étude plus approfondie s'impose donc.

SAP a identifié plusieurs domaines d'application pour le projet. Un exemple est la détection de fraude, qui présente plusieurs challenges relatifs à la non-stationnarité. D'une part, les établissements financiers mettent en place des protocoles de plus en plus sophistiqués pour sécuriser les échanges, comme l'authentification forte de type DSP2. D'autre part, les techniques de fraude évoluent elles-aussi. Enfin le taux de fraude est généralement très faible : collecter suffisamment de cas positifs pour entraîner un modèle nécessite d'utiliser des données couvrant plusieurs années. Tous ces facteurs font que le taux de fraude n'est généralement pas stationnaire à l'échelle d'un jeu de données d'apprentissage, et qu'il est nécessaire d'en tenir compte pour calculer correctement des risques futurs. Un autre challenge est que les fraudes surviennent souvent par salves. Par exemple, un fraudeur va faire un coup d'essai, puis émettre rapidement plusieurs transactions frauduleuses avec des caractéristiques similaires (mode de paiement, localisation, type de produit, ...). Les rares cas positifs peuvent ainsi appartenir à des *clusters*. Il s'agit pour entraîner un modèle prédictif de distinguer les facteurs qui caractérisent un *cluster* unique, et les symptômes qui sont généralisables à un grand nombre de fraudes passées et futures.

## 3 Profil du candidat

Le candidat(e) doit avoir de bonnes notions en mathématique, statistiques et algorithmiques/informatique. Une expérience en traitement de données massives est souhaitable.

Le dossier de candidature en PDF en un seul fichier comportera les éléments suivants :
— CV
— Relevés de notes, M1, M2 (Ing)
— Lettre de motivation
— Lettre(s) de recommandation et/ou des références

Le dossier de candidature est à envoyer par courriel à mlcandidat@gmail.com (en précisant dans le l'objet du mail [CIFRE-SAP]) :

## Références

[Sha88] Lloyd S. Shapley. "A value for n-person games". In : *The Shapley Value : Essays in Honor of Lloyd S. Shapley*. Sous la dir. d'Alvin E.Editor Roth. Cambridge University Press, 1988, p. 31-40. doi : 10.1017/CBO9780511528446.003.

[Fri00] Jerome H. Friedman. "Greedy Function Approximation : A Gradient Boosting Machine". In : *Annals of Statistics* 29 (2000), p. 1189-1232.

[Mas+00] Llew Mason, Jonathan Baxter, Peter Bartlett et Marcus Frean. "Boosting Algorithms as Gradient Descent". In : *In Advances in Neural Information Processing Systems 12*. MIT Press, 2000, p. 512-518.

[CG16] Tianqi Chen et Carlos Guestrin. "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : Association for Computing Machinery, 2016, p. 785-794. isbn : 9781450342322. doi : 10.1145/2939672.2939785. url : https://doi.org/10.1145/2939672.2939785.

[LL17] Scott M. Lundberg et Su-In Lee. "A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems". In : (2017). Sous la dir. d'Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan et Roman Garnett, p. 4765-4774.

[Lun+20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal et Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In : *Nature Machine Intelligence* 2.1 (2020), p. 2522-5839.