# Détection de fraudes bancaires intra-entreprise par *machine learning*

| | |
|---|---|
| **Supérieur hiérarchique :** ***Fraud Management Product Owners*** | **Département :** *PRODUCT* |
| **Lieu** : Saint-Cloud proximité T2, métro ligne 10, gare SNCF | **Durée du stage** : 6 mois |

## Entreprise

Kyriba est un éditeur de logiciel présent en Europe, aux États-Unis et en Asie, qui développe et commercialise en mode SaaS (*Software as a Service*), et à destination des entreprises, des solutions de gestion de trésorerie, de risques financiers (change, taux, …) et de paiements.
Leader sur son marché, la suite logicielle de Kyriba est utilisée par plus de 1600 grands groupes internationaux dans plus de 100 pays. Avec des clients exigeants et prestigieux tels que Amazon, Auchan, Electronic Arts, Qualcomm, Spotify, Uber.

Afin d'assister notre équipe Produit et d'accompagner notre programme de recherche sur ce thème, nous recherchons un ingénieur R&D junior dans le cadre d'un stage de Master.

## Contexte

Les entreprises reçoivent quotidiennement des relevés de compte de leurs nombreux comptes bancaires. Les transactions de débit/crédit qui figurent sur chaque relevé reflètent leur activité économique et comprennent nombre d'informations telles que la méthode de paiement/d'encaissement (chèque, virement, …), la « contrepartie » (le bénéficiaire du paiement ou son origine : fournisseurs, employés, clients, …). Ces données sont classiquement utilisées à des fins de comptabilisation ou de « rapprochement » (identification) avec des prévision. Elles représentent cependant un gisement de données qui peut être utilisé à des fins d'analyse, notamment :

- À des fins de lutte contre la fraude : pour identifier des paiements atypiques (en montant, en fréquence), des encaissements en provenance de sociétés ou d'individus non autorisés, …
- À des fins de conformité, pour identifier des natures de dépense non autorisées sur la base de mots clé prédéfinis.
- À des fins d'identification de transactions spécifiques; exemple: recherche des transactions qui représentent des échéances d'un même emprunt, et qui sont donc similaires à l'exception éventuellement de la date et du montant (mais avec la même référence, description, …).

## Objectifs

Dans le cadre de nouveaux besoins clients, et compte tenu des enjeux pour les entreprises, Kyriba souhaite explorer l'opportunité d'utiliser les méthodes de *machine learning*, appliquées aux données de nos clients, pour détecter plus finement des fraudes potentielles. Le but global de ce module de Fraude est d'alerter les entreprises en cas d'activité anormale (fraude ?, erreur ?) détectée à partir de leurs nouveaux relevés de compte comparés à l'historique des relevés de compte.
L'objet du stage est l'analyse des transactions bancaires lues dans les relevés de compte afin de détecter des transactions anormales. Son objectif est de tester et de *benchmarker* différents algorithmes d'apprentissage, et d'effectuer une analyse de leur pertinence pour une application dans le cadre contraint des données de Kyriba. Les données utilisées dans le cadre de ce projet sont des données anonymisées extraites de bases de données clients.

L'enjeu est d'une part la pertinence des alertes et d'autre part la performance de la détection, compte tenu des volumes élevés de transactions bancaires traités par les clients. Dans un premier temps, il est envisagé d'utiliser un *classifier* dont les paramètres sont appris toutes les nuits. Puis d'évoluer vers un apprentissage *online*, ceci compte tenu de l'émergence de processus de paiement dits instantanés (qui représentent des encaissements instantanés pour les bénéficiaires de paiement).
Les deux principales spécificités des données manipulées par Kyriba pour ce contexte sont que 1. celles-ci sont non-étiquetées – nous n'avons pas de référence d'apprentissage sur ce qu'est une transaction effectivement frauduleuse – , et 2. qu'elles sont plongées dans un espace de grande dimension – et donc clairsemées. Les approches envisagées pour le moment comprennent des approches bayésiennes, des approches de *clustering*, ou de *random forest*. Mais le stagiaire devra aussi être force de proposition afin de proposer d'autres algorithmes en adéquation avec les contraintes du projet.

## Connaissances demandées

- Connaissance des principales méthodes de machine learning, en particulier concernant l'apprentissage non-supervisé.
- Connaissance de Python, Java ou R ; SQL est un plus.
- Connaissance des principales librairies d'apprentissage dans le langage choisi.
- Être capable de communiquer au fil de l'eau avec le reste des équipes sur les avancées et problèmes rencontrés.
- Connaissance suffisante de l'anglais comme langue de communication en sus du français.

Il sera alloué pendant la durée du stage une indemnité, ainsi qu'un badge d'accès au Restaurant Inter-Entreprises/des Tickets Restaurant, le remboursement du titre de transport à 50%, l'accès au CE.