Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTERE DE L'ECONOMIE ET DES FINANCES

Le « data mining », une démarche pour améliorer le ciblage des contrôles

La lutte contre la fraude aux finances publiques a été renforcée ces dernières années et a affiché des résultats en constante progression grâce à la mobilisation des services de contrôle des différentes administrations et organismes de protection sociale. Pour autant, l'ensemble des acteurs de la lutte contre la fraude est confronté à la nécessité d'une constante adaptation des politiques de contrôle et doit faire face à de nouveaux enjeux. Les comportements de fraude s'adaptent en permanence à leur environnement et des fraudes plus complexes, plus sophistiquées se sont développées.

Dans ce contexte, le « data mining » est perçu comme un moyen efficace et innovant permettant à la fois :

- d'améliorer le ciblage des contrôles et ainsi permettre - à moyens constants - d'accroître le nombre de fraudes détectées, redressées et recouvrées,

- de détecter plus rapidement les fraudes, notamment les plus complexes, afin d'en stopper les conséquences financières.

Mener à bien un projet de data mining implique néanmoins le respect d'un certain nombre de conditions de réussite à chacune des étapes du projet (définition des objectifs, mobilisation des compétences, sélection des données à utiliser, choix des modèles statistiques et/ou économétriques, démarches CNIL, expérimentation, évaluation de l'efficacité du modèle, conduite du changement, adaptation des modèles dans le temps…).

Pour répondre à cet objectif, la DNLF a mis en place un groupe de travail interministériel dédié qui associe les administrations financières et les organismes de protection sociale. Ce groupe est le lieu d'échange des bonnes pratiques et des écueils à éviter pour la mise en œuvre d'un projet de data mining. Plusieurs expérimentations sont actuellement menées au sein des administrations financières et des organismes de protection sociale. La DNLF accompagne chacun des acteurs de la lutte contre la fraude.

Les pouvoirs publics ont récemment encore réaffirmé toute l'importance de la lutte contre la fraude aux finances publiques. Les différentes administrations et organismes de protection sociale sont confrontés à de nouveaux enjeux : les comportements de fraude s'adaptent en permanence à leur environnement et des fraudes plus complexes, plus sophistiquées se sont développées. La lutte contre la fraude aux finances publiques exige donc des réponses adaptées à la diversité des situations et des publics concernés. Cela implique la mise en place d'un pilotage stratégique des actions à mener.

Dans ce contexte, la DNLF s'est vue confier la mission de coordination et d'accompagnement des acteurs de la lutte contre la fraude dans le développement de démarches statistiques visant à renforcer l'efficacité des stratégies de ciblage des contrôles. Le data mining constitue une démarche innovante, reposant sur des concepts et des outils qu'il convient de définir précisément avant d'en exposer les différentes étapes.

1. Définition du « data mining »

De manière générale, le data mining est une démarche méthodologique rigoureuse développée en vue de révéler de l'information contenue dans les systèmes d'information, en mettant en exergue d'éventuelles corrélations significatives entre les données observées. Cette démarche globale est composée d'un ensemble de techniques relevant du domaine des statistiques et des mathématiques permettant, à partir d'un important volume de données, d'extraire des informations visant à améliorer la connaissance des phénomènes étudiés (dans notre cas, les comportements de fraude) et permettre d'engager des actions adaptées à l'objectif poursuivi (ici, la lutte contre la fraude).

Sur le plan opérationnel, le data mining fait appel à différentes techniques plus ou moins sophistiquées, chacune ayant une finalité propre et des conditions nécessaires à sa mise en œuvre. Ces techniques sont fortement conditionnées par la nature et la qualité des données disponibles dans les systèmes d'information utilisés.

Dans le contexte, deux cas de figure se présentent. Le premier concerne la recherche de critères discriminants de fraude en l'absence d'historique de cas détectés. Le second cas de figure se rapporte à un contexte pour lequel on dispose d'informations relatives à des cas de fraude déjà identifiés. L'analyse des situations de fraude en réseau implique, quant à elle, le recours à des techniques particulières (recherche d'informations inter-reliées).

Dans la suite de la note, nous utiliserons le terme générique de « *fraude* » pour qualifier toutes les situations conduisant à ne pas déclarer ce qui devrait l'être.

2. A quoi sert le data mining

Cette approche permet de dépasser la simple observation de données brutes et vise à mettre en exergue l'impact réel de chaque paramètre sur le phénomène étudié (*i.e.* la fraude), tout en tenant compte des éventuelles corrélations trompeuses. L'amélioration de la connaissance des comportements de fraude doit ensuite permettre la mise en place d'une action efficace, *i.e.* permettre la détection le plus en amont possible et renforcer l'efficacité des stratégies de ciblage.

Le recours à une telle démarche représente une opportunité pour optimiser les stratégies de ciblage habituelles dans la mesure où, dans certains cas, les acteurs de la lutte contre la fraude ont indiqué que les contrôles aléatoires se sont révélés tout aussi efficaces que les contrôles ciblés.

Des exemples réussis de data mining à l'étranger font ressortir un retour sur investissement élevé.

3. Différentes étapes d'une démarche de data mining

Le data mining ne se limite pas aux seuls outils statistiques et informatiques. Cela implique, par conséquent, la mise en place d'un pilotage stratégique et un suivi des différentes étapes du projet dans le respect de conditions indispensables à sa réussite. Il convient, en effet, de respecter un enchaînement scrupuleux d'étapes qui conditionnent la réussite du projet.

3.1. Définition de la cible à atteindre

La première étape d'un projet de data mining consiste à bien préciser l'objectif poursuivi, notamment en termes d'indicateurs à définir de façon précise et à maximiser. Ces indicateurs serviront également au suivi et à l'évaluation de l'efficacité de la démarche.

Il peut s'agir, par exemple, de la probabilité de procéder à un redressement lors d'un contrôle, du montant du redressement en valeur absolue, du montant de redressement rapporté au montant des cotisations ou de l'impôt dû, du nombre d'allocataires détectés en situation de fraude, du montant de prestations indues stoppées avant versement…

La définition et la construction de ces indicateurs est déterminante pour la conduite du projet de data mining car ceux-ci serviront à la construction du modèle. Ces indicateurs correspondent à la (aux) variable(s) à expliquer en fonction de facteurs d'influence à tester.

3.2. Définition du périmètre

La définition du périmètre est fondamentale, tant du point de vue du champ d'investigation (secteur d'activité, zone géographique, catégorie d'impôt, segment de cotisants, type d'allocataire…) que du point de vue de la caractérisation de la fraude.

S'agissant du champ d'investigation, tout projet de data mining - pour être efficace – doit bien délimiter la population à examiner pour être en mesure d'identifier les facteurs d'influence de la fraude pour cette population et ainsi sélectionner les individus statistiques les plus à risque.

En effet, l'hétérogénéité des comportements de fraude est d'autant plus importante que l'on raisonne sur des populations différentes. Il est dès lors plus délicat de cerner les critères de risque qui soient valables pour l'ensemble de la population sur des périmètres flous ou trop vastes.

De même, l'hétérogénéité des pratiques de contrôle sur le territoire est également de nature à perturber l'analyse et à compromettre l'identification de critères de ciblage. C'est pourquoi, il est conseillé de délimiter le périmètre de l'expérimentation à un domaine particulier et une catégorie spécifique d'individus.

S'agissant de la caractérisation de la fraude, compte tenu des limites de certains systèmes d'information, il ne sera pas possible de distinguer les cas de fraude des cas de non-respect des obligations déclaratives.

3.3. Fiabilisation des données

L'identification des besoins en termes de données et la fiabilisation de celles-ci sont ensuite indispensables pour la constitution de ce qu'il est convenu d'appeler « entrepôt de données ». Ce dernier étant destiné à recueillir l'ensemble des informations sur lequel vont porter les analyses. Ainsi, un examen approfondi des données disponibles dans le système d'information devra être mené de façon à connaître celles qui sont exploitables, celles qui peuvent être redressées et celles qui ne sont pas suffisamment fiables pour un traitement statistique. Cela suppose notamment de parfaitement maîtriser le(s) circuit(s) d'enregistrement et de stockage des données. En effet, exploiter l'historique du contrôle des individus suppose que les données ne soient pas écrasées par des traitements informatiques réguliers. De même, il conviendra de s'assurer qu'un processus d'agrégation des données ne compromette pas l'exploitation de données fines, indispensables au projet de data mining, notamment en termes de types de contrôle ou de motifs de notification de redressement. Il conviendra enfin de vérifier que la part de valeurs manquantes dans les données est faible et que celles-ci peuvent être corrigées.

Une fois l'entrepôt de données constitué et afin de mieux connaître la population sur laquelle le data mining va porter, une première analyse descriptive des caractéristiques de cette population permettra de mieux appréhender les possibles critères de risque de fraude. Ces derniers serviront à la constitution de modèles économétriques permettant de tester leur validité. Une fois connues les informations exploitables et une fois identifiés les critères supposés pertinents pour le ciblage, il conviendra de construire les variables utiles au(x) traitement(s) statistique(s).

3.4. Déclaration auprès de la CNIL

En parallèle, la finalité de l'exploitation des données pour le projet de data mining doit faire l'objet d'une démarche auprès de la CNIL, si l'utilisation des informations à des fins d'amélioration du ciblage des contrôles est différente de celle qui prévalait antérieurement.

Des autorisations ont d'ores et déjà été délivrées par la CNIL. Plusieurs dossiers ont également été déposés ou sont en cours d'élaboration au sein des administrations et organismes en charge de la lutte contre la lutte.

3.5. Identification des compétences ressources à mobiliser

L'efficacité d'une démarche de data mining est fortement conditionnée par la qualité des données collectées, mais elle dépend également grandement des contrôles réalisés et plus particulièrement de l'adhésion des corps de contrôle à cette démarche. Un effort de pédagogie doit donc être réalisé de façon à ce que les directions et les équipes en charge du contrôle ne se sentent pas dépossédées de leurs missions.

Les techniques statistiques utilisées ne remplaceront pas les contrôles et l'expertise des corps de contrôle pour la détection de phénomènes de fraude. C'est pourquoi, il est crucial

de mobiliser l'ensemble des compétences métier - tant au niveau du pilotage que de la réalisation des contrôles - et de les associer très en amont de la démarche de façon à assurer leur adhésion complète. Les compétences statistiques sont naturellement essentielles pour la conduite de ce projet, de même que celles dédiées aux systèmes d'information. Mais, les résultats des traitements statistiques doivent être soumis aux responsables métiers pour validation et/ou tests ultérieurs.

Afin de mieux associer les directions en charge du contrôle, des statistiques et des systèmes d'information, il peut être opportun de créer une équipe-projet dédiée en interne. Cette équipe serait composée de l'ensemble des compétences en matière de contrôle, de statistiques et d'informatique nécessaires et se verrait assigner des objectifs propres. Un dialogue régulier devrait, par ailleurs, être instauré au sein de l'équipe de façon à garantir la plus grande efficacité possible dans le déroulement du projet.

3.6. Réalisation des contrôles et évaluation de l'efficacité de la démarche

L'efficacité du data mining dépend de l'exhaustivité des contrôles menés. Seule cette condition permet d'éviter que l'analyse des résultats des contrôles ne soit biaisée.

Il est important, en second lieu, de définir dès que possible en amont des indicateurs (*cf. paragraphe 3.1*). Les éléments constituants ces indicateurs pourront permettre une comparaison avec les résultats obtenus habituellement ou bien avec ceux des contrôles aléatoires, qui doivent être en nombre suffisant pour alimenter la veille sur de nouvelles fraudes.

Les résultats issus des contrôles ciblés par data mining devront ensuite être intégrés à l'entrepôt de données servant à l'élaboration des plans de contrôle ciblés, de façon à actualiser le modèle et éventuellement tester de nouveaux critères de risque de fraude.