



Centre National d'Appui au Datamining



## **LE MODÈLE DATAMINING DONNEES ENTRANTES 2018**

### **DOCUMENTATION TECHNIQUE**

*septembre 2020*

## Table des matières

---

1. Objet.....	3
2. Source.....	3
3. Définition de la cible.....	3
4. Construction de la table d'apprentissage.....	4
5. Travaux de « Data management » .....	4
6. Modélisation du risque d'indu long .....	5
7. Estimation par validation croisée des performances théoriques du modèle .....	5
8. Scoring et déploiement .....	5
9. Résultats du modèle en production .....	6
10. Travaux complémentaires.....	7
11. Tableau des variables du modèle DMDE 2018.....	8

## 1. Objet

Le modèle Datamining Données Entrantes (DM DE) a pour objet de prioriser les dossiers allocataires afin d'orienter les contrôles sur place ou sur pièce vers ceux qui sont le plus à risque d'indus. Le modèle précédemment en production avait été développé par la DSER et les statisticiens girondins en 2013.

Les enjeux d'une nouvelle modélisation sont :

- Proposer un modèle plus actuel, prenant en compte à la fois les évolutions législatives, de gestion, les évolutions du SI mais également les changements de comportement des allocataires
- Disposer d'une cible plus adaptée à la lutte contre la fraude (Note DSER 2017-131 « Vers une évolution des modèles datamining Données entrantes pour détecter davantage de fraude »)
- Construire un modèle plus robuste aux évolutions réglementaires à venir (Note Cnad : « Evaluation de la résistance du modèle DMDE aux chocs législatifs »)
- Intégrer des spécificités locales dans le modèle par l'introduction de variables d'environnement socio-économique.

## 2. Source

La réalisation de ce modèle s'appuie sur les données de l'enquête Paiement à bon droit fraudes (PBDF). Afin d'augmenter le nombre d'observations et améliorer la qualité de la modélisation, deux millésimes de l'enquête ont été regroupées (2016 et 2017), ce qui complexifie considérablement les étapes d'enrichissement de la table d'apprentissage, mais apporte beaucoup de robustesse.

La table ainsi constituée compte 14 087 individus et quelques 4360 variables brutes après enrichissement (avant traitement et création d'indicateurs).

## 3. Définition de la cible

Un travail de définition de la cible a été entrepris afin de répondre à la commande de mieux cibler les indus importants (notamment frauduleux) tout en ne s'écartant pas trop de la cible actuellement en production.

Enquêtes 2016 et 2017					
	Nombre d'observations	En % du total	Part des indus de la cible qualifiés de frauduleux (en montants, données pondérées)	Part des indus frauduleux qui sont dans la cible (en montants, données pondérées)	Impact financier (indus + rappels) moyen
Cible 1 indus qualifiés de frauduleux	713	5,0%	100%	100%	7 008 €
Cible 2 indus > 200 euros et d'une durée > à 3 mois (cible actuelle)	2 222	15,6%	59%	100%	4 021 €
Cible 3 indus > 400 euros et d'une durée > à 5 mois	1 833	12,9%	61%	99%	4 681 €
Cible 4 indus > 600 euros et d'une durée > à 6 mois	1 546	10,9%	62%	98%	5 320 €
Cible 5 indus > 1000 euros et d'une durée > à 6 mois	1 347	9,5%	63%	97%	5 926 €
Cible 6 indus > 1500 euros et d'une durée > à 6 mois	1 096	7,7%	66%	96%	6 908 €
Cible 7 indus mensuel moyen > 10 euros	3 350	23,5%	57%	100%	2 854 €
Cible 8 indus mensuel moyen > 30 euros	2 896	20,4%	58%	100%	3 238 €
Cible 9 indus mensuel moyen > 50 euros	2 526	17,8%	58%	99%	3 620 €
Cible 10 indus mensuel moyen > 100 euros	1 852	13,0%	61%	97%	4 559 €
Cible 11 indus mensuel moyen > 200 euros	1 057	7,4%	68%	88%	6 360 €
Total des observations	14 227	100%	-	-	-

La cible 4 a été retenue par le COPIL stratégique Datamining de la branche Famille, soit « indus nets supérieurs à 600€ avec durée supérieure à 6 mois ». Cela permet de couvrir 98% des indus frauduleux des deux enquêtes PBDf tout en conservant une cible ni trop rare ni d'occurrence trop fréquente.

#### 4. Construction de la table d'apprentissage

Cette étape consiste à récupérer un maximum de données dans le système d'information de la branche famille. Ces données proviennent des tables :

- Allstat/FR1 : données et historique de Cristal (système de gestion des dossiers allocataires)
- Allstat/FFAI : données et historique des faits générateurs Cristal
- GCA : données et historique sur la Gestion du Contact Allocataire
- Telconta : données et historique sur les contacts téléphoniques
- SDP : données et historique sur les pièces
- Web : données et historique sur les connexions au Caf.fr
- Corali : données issues du SI Contentieux et recours amiable
- SIAS : données issues du SI Action sociale (Aides financières aux familles)
- Cristal : données de l'infocentre Cristal

Elles sont recueillies sur un historique de 18 mois maximum.

Une FEB a été rédigée afin de pouvoir faire exécuter par la DSI, le programme de constitution de la table d'apprentissage dans les entrepôts locaux de chaque Caf.

Ont également été intégrées des variables locales d'environnement socio-économique issues de la BCA (base communale allocataires) et hors SID : données sur les allocataires à bas revenus et fortement dépendants des prestations, sur les taux d'effort ; données INSEE sur le logement et sur les taux de chômage localisé.

La table d'apprentissage est constituée, après un premier traitement, de 1 700 variables et indicateurs.

Il s'agit de variables concernant la situation de l'allocataire à la veille du contrôle mais également des variables résumant l'historique de ces situations sur 3, 6, 12 ou 18 mois.

#### 5. Travaux de « Data management »

Cette étape consiste à vérifier la qualité des données et les préparer en vue de la modélisation. Elle est d'autant plus importante compte tenu du grand nombre de variables recueillies dans la table d'apprentissage.

- Sélection des variables : suppression des variables contenant trop peu d'observations et des variables pour lesquelles le lien avec la cible est trop faible
- Recodage des variables alphanumériques et discrétisation des variables quantitatives,
- Calcul des corrélations 2 à 2 des variables et réalisation de CAH pour identifier les variables très corrélées, c'est-à-dire portant la même information

Des recodages sont à nouveau effectués lors de la phase de modélisation pour rendre certaines variables plus pertinentes.

## 6. Modélisation du risque d'indu long

- Mise en œuvre d'un bootstrap (300 itérations) pour sélectionner les variables significatives sur au moins 50% des modèles estimés par régression logistique.
- Parmi les variables sélectionnées, recherche et identification des interactions apportant un gain de performance significatif au modèle.
- Les performances théoriques du modèle sont estimées par validation croisée et bootstrap.
- Le modèle sélectionné est ré-estimé sur la table d'apprentissage entière pour calculer les coefficients.

## 7. Estimation par validation croisée des performances théoriques du modèle

<b>Aire sous la courbe LIFT</b>	<b>60,0%</b>
<b>Aire sous la courbe ROC</b>	<b>80,0%</b>
<b>% concordant</b>	<b>79,8%</b>
<hr/>	
<b>Statistiques sur les 5% les plus scorés</b>	
<b>% cibles</b>	<b>38,2%</b>
<b>% dossiers frauduleux</b>	<b>20,1%</b>
<b>% dossiers avec IF</b>	<b>68,6%</b>
<b>% dossiers avec indus</b>	<b>64,6%</b>
<b>Montant moyen IF</b>	<b>2 604 €</b>

## 8. Scoring et déploiement

- Ecriture sous SAS d'un programme de récupération des variables retenues dans le modèle final.
- Ecriture de l'algorithme calculant la probabilité d'indu long à partir de la valeur des paramètres recueillis lors de la modélisation logistique.
- Le modèle a été testé dans une vingtaine de Caf à compter de février 2019 et déployé dans l'ensemble des caisses en décembre 2019.

## 9. Résultats du modèle en production

La DSER assure un suivi des résultats des contrôles lancés via le modèle DMDE. Elle compare les résultats des contrôles lancés à partir de février 2019 (début des tests) et clos depuis entre les Caf ayant testé le modèle depuis février 2019 et les autres Caf.

Pour **les contrôles sur place**, le tableau récapitulatif :

Contrôles sur place lancés à partir de février 2019, clos entre février 2019 et juillet 2020	Nombre de contrôles	Mt Indus bruts total (en millions d'euros)	Mt moyen indus - Nb de contrôles (en euros)	Mt moyen indus - Nb d'indus (en euros)	% de contrôles avec indu brut	% de contrôles avec rappel brut	% de contrôles avec indu net	% de contrôles avec rappel net	% de contrôles avec indu brut ou rappel brut	% de contrôles avec indu cible DMDE 2014	% de contrôles avec indu cible DMDE 2018
Tous	106 978	128,3	1 199	2 093	57,3	47,8	43,5	22,1	65,6	28,7	25,5
Groupe Témoin	76 418	85,4	1 117	1 990	56,1	48,0	42,0	22,8	64,9	27,1	24,2
Groupe Test	30 560	42,9	1 403	2 333	60,2	47,5	47,3	20,2	67,5	32,5	28,8

  

Contrôles sur place lancés à partir de décembre 2017, clos entre décembre 2017 et décembre 2018	Nombre de contrôles	Mt Indus bruts total (en millions d'euros)	Mt moyen indus - Nb de contrôles (en euros)	Mt moyen indus - Nb d'indus (en euros)	% de contrôles avec indu brut	% de contrôles avec rappel brut	% de contrôles avec indu net	% de contrôles avec rappel net	% de contrôles avec indu brut ou rappel brut	% de contrôles avec indu cible DMDE 2014	% de contrôles avec indu cible DMDE 2018
Tous	90 228	93,02	1 031	1 914	53,9	44,2	41,6	20,5	62,1	26,4	22,9
Groupe Témoin	65 371	64,4	986	1 822	54,1	44,6	41,6	20,7	62,4	26,1	22,5
Groupe Test	24 857	28,6	1 150	2 161	53,2	43,2	41,6	19,8	61,4	27,2	23,8

Pour **les contrôles sur pièces**, le tableau récapitulatif :

Contrôles sur pièces lancés à partir de février 2019, clos entre février 2019 et juillet 2020	Nombre de contrôles	Mt Indus bruts total (en millions d'euros)	Mt moyen indus - Nb de contrôles (en euros)	Mt moyen indus - Nb d'indus (en euros)	% de contrôles avec indu brut	% de contrôles avec rappel brut	% de contrôles avec indu net	% de contrôles avec rappel net	% de contrôles avec indu brut ou rappel brut	% de contrôles avec indu cible DMDE 2014	% de contrôles avec indu cible DMDE 2018
Tous	265 065	79,1	299	1 223	24,4	20,0	19,4	11,2	30,7	10,6	8,1
Groupe Témoin	184 871	50,9	275	1 164	23,7	19,8	18,8	11,4	30,2	9,9	7,5
Groupe Test	80 194	28,2	352	1 347	26,1	20,5	21,0	10,8	31,8	12,2	9,6

  

Contrôles sur pièces lancés à partir de décembre 2017, clos entre décembre 2017 et décembre 2018	Nombre de contrôles	Mt Indus bruts total (en millions d'euros)	Mt moyen indus - Nb de contrôles (en euros)	Mt moyen indus - Nb d'indus (en euros)	% de contrôles avec indu brut	% de contrôles avec rappel brut	% de contrôles avec indu net	% de contrôles avec rappel net	% de contrôles avec indu brut ou rappel brut	% de contrôles avec indu cible DMDE 2014	% de contrôles avec indu cible DMDE 2018
Tous	240 178	55,53	231	1 117	20,7	17,3	16,6	10,4	27,0	8,4	6,3
Groupe Témoin	176 302	40,2	228	1 091	20,9	17,7	16,7	10,6	27,4	8,4	6,2
Groupe Test	63 876	15,3	240	1 192	20,1	16,3	16,4	9,6	26,0	8,5	6,4