

Contrôle des déclarations de ressources des demandes de Complémentaire Santé Solidaire (C2S) sans Contribution Financière (CF)

Datamining – 2^{ème} modèle

Datamining C2S - 2ème modèle

❑ Contexte

2015 : lancement du contrôle CMUc basé essentiellement sur un tirage aléatoire de dossiers

2018 : généralisation du 1^{er} modèle de Datamining CMUc avec un objectif annuel par caisse de 50% de dossiers ciblés Datamining/ciblages locaux et 50% de tirage aléatoire/signalements

2020 : élaboration d'un 2ème modèle de Datamining C2S sans CF **afin de conforter la première modélisation**

❑ Objectifs et Finalités du Datamining

Rendre le contrôle C2S **plus efficient** :

- Des contrôles **mieux ciblés sur les risques d'anomalies et de fraude**
- **Moins de dossiers** pour optimiser les ETP

Rappel : contrôle long avec droit de communication bancaire, jusqu'à 6 mois d'investigations nécessaires pour les dossiers avec anomalies

Datamining C2S - 2ème modèle

Une volumétrie moindre pour de meilleurs résultats

✓ Résultats globaux du programme national pérenne

2016			2017			2018			2019		
Nb dossiers contrôlés	Taux d'ano.	Taux de Fraude	Nb dossiers contrôlés	Taux d'ano.	Taux de Fraude	Nb dossiers contrôlés	Taux d'ano.	Taux de Fraude	Nb dossiers contrôlés	Taux d'ano.	Taux de Fraude
62 460	19,46%	7,82%	41 337	17,95%	7,02%	24 904	23,89%	9,43%	19 128	23,73%	10,24%

Source: Sofi Web

✓ Montants de préjudice par année de reporting 2016-2019

2016			2017			2018			2019		
Nb dossiers av. préjudice	Subi	Evité	Nb dossiers av. préjudice	Subi	Evité	Nb dossiers av. préjudice	Subi	Evité	Nb dossiers av. préjudice	Subi	Evité
5 990	826 315 €	3 501 552 €	7 955	1 655 409 €	5 519 174 €	6 010	1 435 101 €	4 265 906 €	4 437	1 339 228 €	3 312 329 €
Total préjudices	4 327 867 €		Total préjudices	7 174 583 €		Total préjudices	5 701 007 €		Total préjudices	4 651 556 €	
Montant moyen / dossier	723 €		Montant moyen / dossier	902 €		Montant moyen / dossier	949 €		Montant moyen / dossier	1 048 €	

Datamining C2S - 2ème modèle

2^{ème} modèle – MODÈLE GÉNÉRAL

- ❑ **Données et variables** : uniquement tirage aléatoire et signalements

Données extraites de l'outil SOFI Web sur la période **01.05.2018** au **31.12.2019**

11 184 dossiers C2S sans CF avec un **taux global d'anomalies de 19%**

VARIABLES	DESCRIPTIF (Les classes soulignées sont utilisées comme référence dans le modèle)

Datamining C2S - 2ème modèle

2^{ème} modèle – MODÈLE GÉNÉRAL

❑ Résultats

VARIABLES	CLASSE DE REFERENCE	ODDS-RATIO	Intervalle de Confiance – Borne inf.	Intervalle de Confiance – Borne sup.

Le score de risque d'anomalies sera plus élevé pour les dossiers de

Datamining C2S - 2ème modèle

2^{ème} modèle – MODÈLE RESTREINT



Données des comptes bancaires issues d'une interrogation manuelle de FICOBA par les caisses => **récupération de cette donnée impossible** dans le cas d'une validation et généralisation du modèle

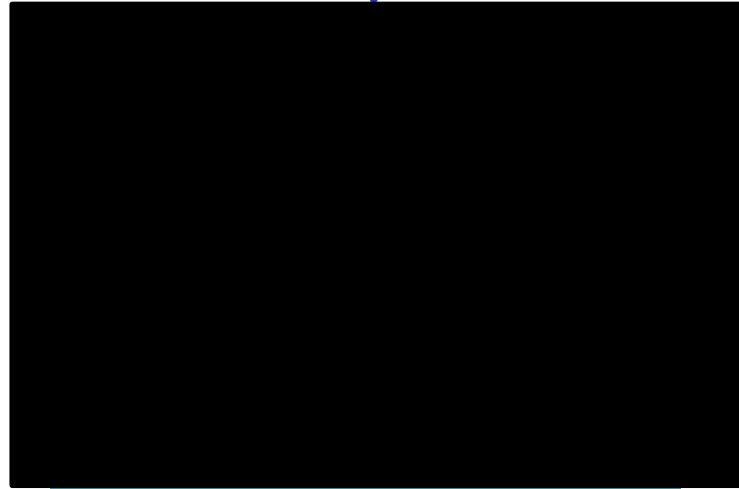
☐ Résultats du modèle restreint (*hors donnée compte bancaire*)

VARIABLES	CLASSE DE REFERENCE	ODDS-RATIO	Intervalle de Confiance – Borne inf.	Intervalle de Confiance – Borne sup.
-----------	---------------------	------------	--	--

[Redacted content]				
--------------------	--	--	--	--

Datamining C2S - 2ème modèle

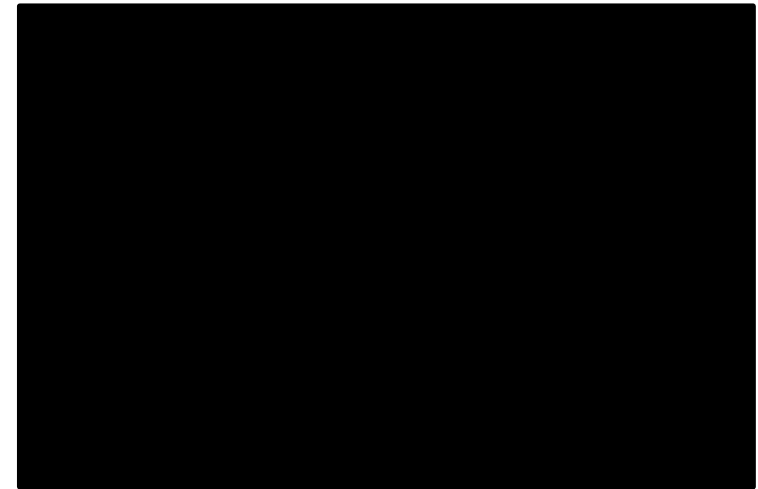
Comparaison entre le 1^{er} et le 2^{ème} modèle



Tirage aléatoire :
Taux d'anomalies = **18.9%**

Datamining (modèle restreint) :
Taux d'anomalies théorique⁽¹⁾ = **28.5%**

Datamining
(modèle général à 16 variables) :
Taux d'anomalies théorique = **33.1%**



Tirage aléatoire :
Taux d'anomalies = **19%**

Datamining (modèle restreint) :
Taux d'anomalies théorique = **28.8%**

Datamining
(modèle général à 7 variables) :
Taux d'anomalies théorique = **33.5%**

Datamining C2S - 2ème modèle

□ Perspectives

Deux options pour améliorer l'efficacité du ciblage par Datamining en passant à un modèle moins contraint :

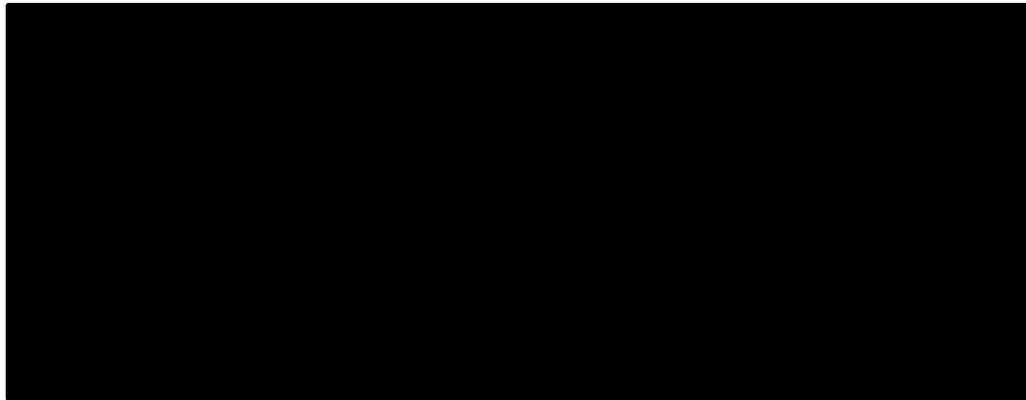
1. Rendre le **croisement possible** entre **Base Ressources et SIAM Erasme >> INDIGO ?**
2. **Pouvoir disposer de la donnée *Nombre de comptes bancaires*** (contenu dans la base de données FICOBA) dans le cadre des **travaux d'échanges de données avec les organismes de protection sociale** (suite COPIL fraude DSS - octobre 2020)

ANNEXES

Datamining C2S – 1^{er} modèle

❑ Méthode

- ✓ **Données** relatives aux dossiers contrôlés, remontées par les CPAM
- ✓ **Variable à prédire** : Anomalie **OUI / NON** avec modalité OUI = regroupement des 4 modalités suivantes :



- ✓ **Modèle retenu** : régression logistique (modèle LOGIT)

Datamining C2S – 1er modèle

Rappel sur le 1^{er} modèle – MODÈLE GÉNÉRAL

Au départ, **42 000 foyers** CMUc* avec **900 variables** issues de sources différentes (SOFI Web, FICOBA, SIAM Erasme)



Sélection des 16 variables les plus discriminantes dont **14 significatives**

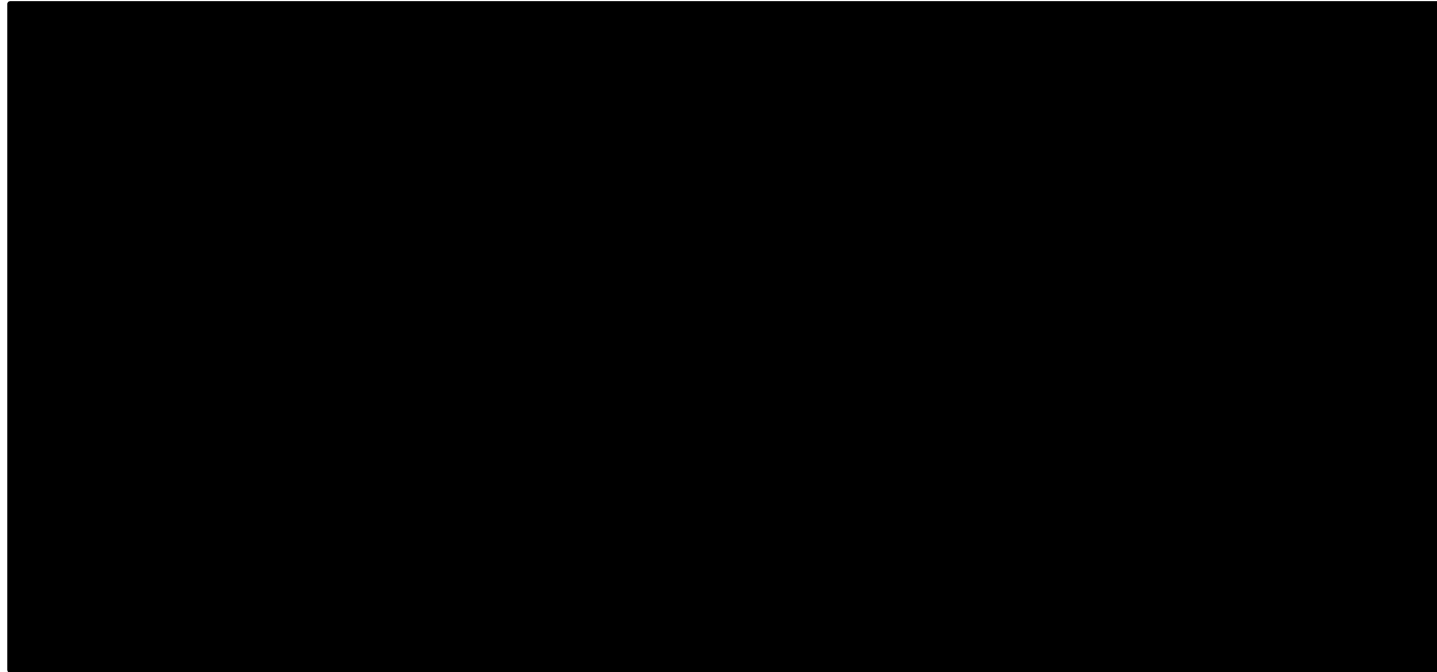
* Janv. 2015 à Dec. 2016

Effet	Odds Ratio	Borne inférieure IC95%	Borne supérieure IC95%

Datamining C2S – 1er modèle

Rappel sur le 1^{er} modèle – MODÈLE GÉNÉRAL

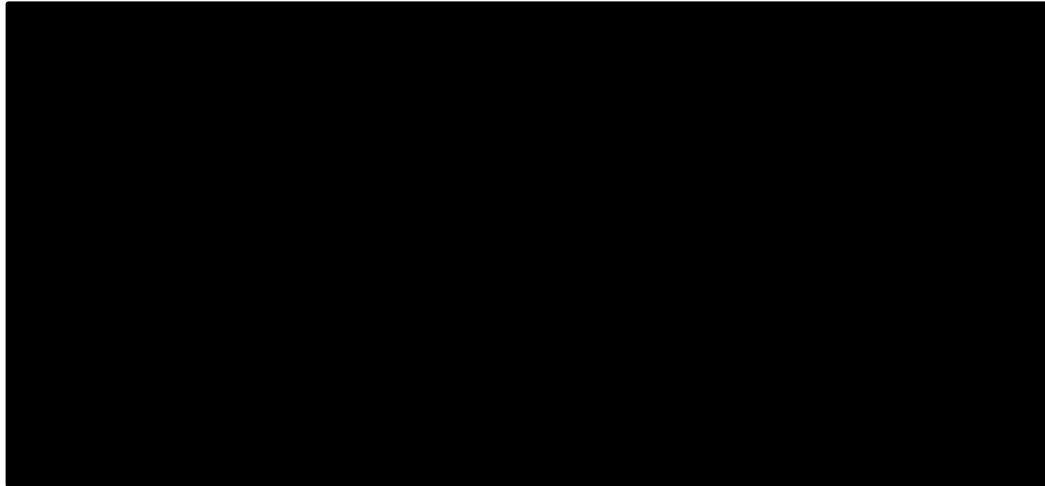
Dans ce modèle général, on observe **8 variables** sur 14 qui augmentent significativement le risque d'être en anomalies :



Datamining C2S – 1er modèle

Rappel sur le 1^{er} modèle – MODÈLE GÉNÉRAL

A contrario, **6 variables** sur 14 diminuent significativement le risque d'être en anomalies :



Datamining C2S – 1er modèle

Rappel sur le 1^{er} modèle – MODÈLE RESTREINT



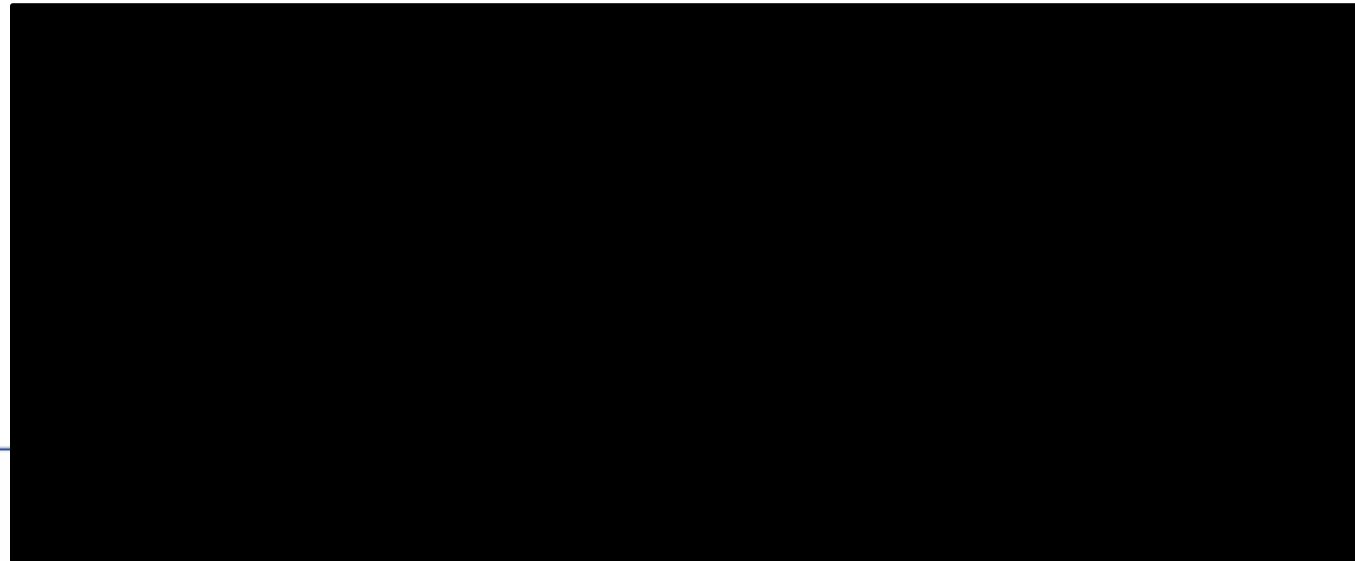
PROBLÉMATIQUE DU MODELE GENERAL

Validation et généralisation de ce premier modèle général **impossible au niveau CPAM** car nécessite notamment un **croisement de données non autorisé** entre Base Ressources et SIAM Erasme régional 2018

➤ ALTERNATIVE

Ajuster un modèle plus restreint **en se limitant aux données disponibles dans la Base Ressources >> 5 variables significatives**

Score de risque
d'anomalies plus élevé
pour les dossiers suivants :



Datamining C2S – 1er modèle

Rappel sur le 1^{er} modèle – MODÈLE RESTREINT

❑ Comparaison des méthodes de ciblage

Pour comparer les propositions, on détermine le taux d'anomalies théorique après ciblage :

- ✓ On ajuste le modèle sur 70% de notre base d'apprentissage et on simule un ciblage sur les 30% restant (= échantillon de validation)
- ✓ On calcule la prédiction d'anomalies pour l'échantillon de validation puis on calcule le taux d'anomalies pour les 10% ayant un score de risque le plus élevé

- Tirage aléatoire : Taux d'anomalies = **18.9%**
- Datamining (modèle général) : Taux d'anomalies théorique = **33.1%**
- Datamining (modèle restreint) : Taux d'anomalies théorique = **28.5%**