

# Méthodes de détection de fraude pour les assurances

Présenté par **René Portillo**  
Encadrant d'entreprise **M. Moisés Rodriguez**  
Encadrant académique **M. Christophe Prud'homme**

**Stage de Master 2 – Calcul Scientifique Mathématiques de l'Innovation**

**Université** de Strasbourg

UFR de mathématique et d'informatique

Université de Strasbourg

**SHIFT**

# Sommaire

**SHIFT**

## **Introduction**

- a) Présentation de l'organisme d'accueil
- b) Objectifs
- c) Mapping, détection et scénarios

## **I – Détection de fraude via des scénarios classiques**

- a) Architecture de la solution
- b) Exemples

## **II - Détection de fraude via Machine Learning**

- a) Présentation des données
- b) LightGBM et architecture
- c) Résultats

# L'assurance

SHIFT

## Property and Casualty (P&C)

### Property :

- Résidence principale
- Voiture
- Les bureaux

### Casualty (Responsabilité civil):

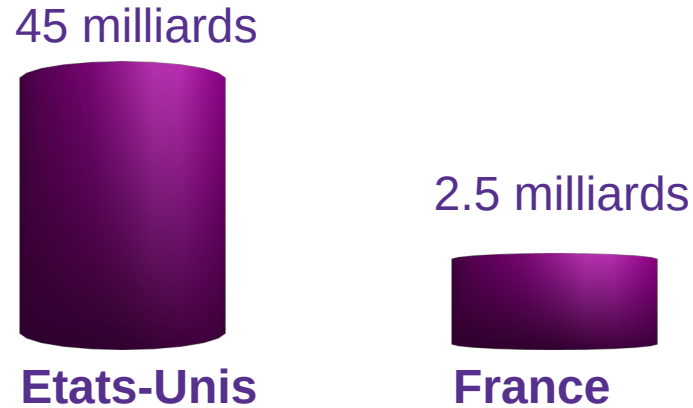
- Dommages dans un accident à un tiers
- Faute médicale

## Health & Life

- Une assurance contre un certain groupe de maladie
- Une assurance pour le travail
- Une assurance vie

# La fraude à l'assurance

SHIFT



Deux types de fraudes :

**Fraude Dure** : Invention complète d'un sinistre (ex: faux accident, incendie volontaire).

**Fraude Simple** : Exagération des dommages d'un sinistre réel pour inclure des dégâts antérieurs.

# Présentation de l'organisme d'accueil

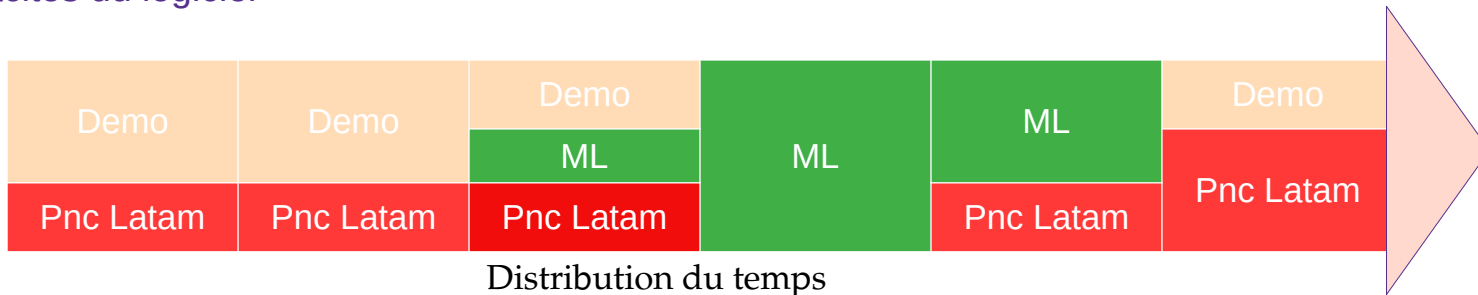
- **Shift Technology** est une entreprise française **fondée en 2014** qui a pour but de détecter la fraude via l'intelligence artificielle
- Elle compte avec plus de **500** salariés
- Elle a des bureaux à Paris, Boston, Mexico, Japon, Singapour, Londres, Madrid et Brésil
- J'ai intégré l'entreprise en tant que **data scientist**



**Liberty**  
**Insurance™**

# Objectifs du stage

- ▶ **Appuyer mon encadrant avec la gestion de trois clients d'amérique latine PnC**
  - Amélioration constante des algorithmes mis en production
  - Création de nouveaux algorithmes pour la détection de fraude
  - Présentation hebdomadaires avec les clients
- ▶ **Mettre en production un modèle de Machine Learning pour la détection de fraude matérielle et corporelle.**
  - Calibration d'un modèle de machine learning pour un assureur espagnol pour des dommages corporels
  - Participation dans le build d'un deuxième client espagnol, conception des deux modèles de machine learning
    - Génération du dataset, nettoyage des données, entraînement du modèle, mise en production
- ▶ **Développer des cas de démonstration pour les équipes commerciales.**
  - Création ou adaptation de 15 cas démos pour l'équipe de Pre Sales et GoToMarket de sorte à montrer les capacités du logiciel



Distribution du temps

# Outils et méthodologies

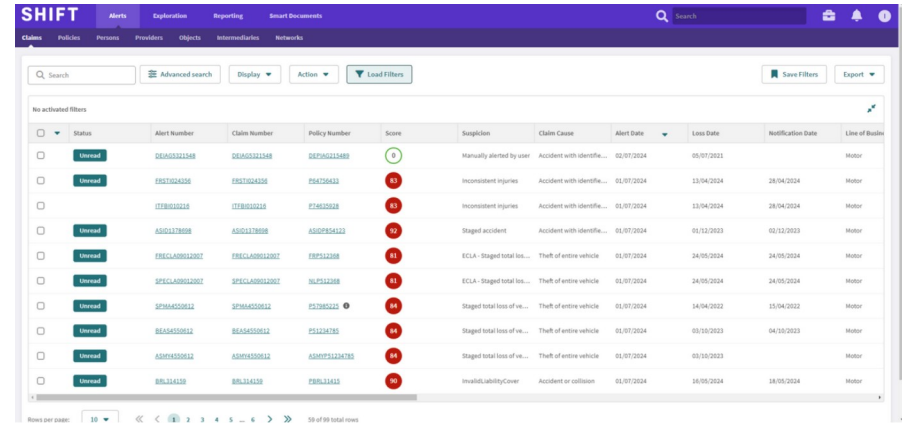
- ▶ **Pour la calibration et création scénarios; comme pour la création des cas de démonstration**
  - Le langage de programmation était le **C#**, avec l'utilisation de **LINQ**
  - **SQLServer** pour la gestion des bases de données
- ▶ **Pour la partie Machine Learning**
  - **Python**, en utilisant les librairies pandas, pandas-profiling, numpy, matplotlib, et LightGBM
- ▶ **Pour l'intégration continue et la gestion de version**
  - La **CI/CD** était gérée avec **TeamCity** pour la génération des builds et la gestion des tests unitaires et **Octopus** pour le déploiement dans les différents tenants
  - **RDC (Remote Desktop Connection)** pour l'accès au serveurs de production, préproduction et staging à distance
  - **Git et GitExtesions** pour la gestion de version

De façon générale:

- ▶ Excel pour le traitement des fichiers .csv et des manipulations simples des résultats
- ▶ **Jira** pour la gestion des tickets et **Confluence** pour la documentation

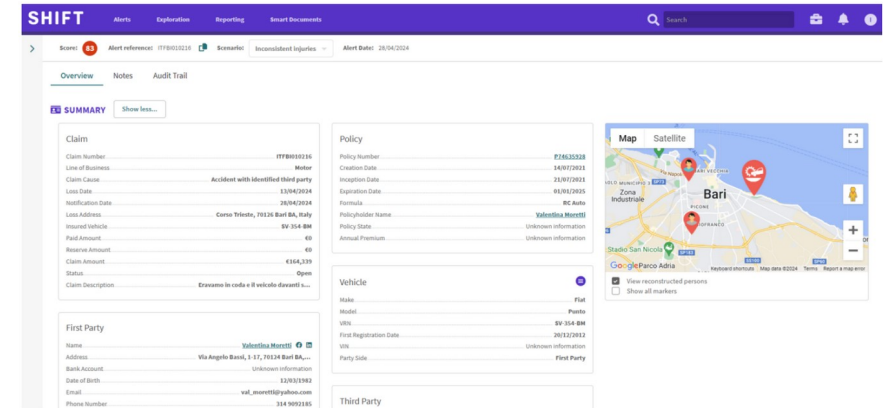
# Force

- **Force** est le logiciel de détection de fraudes et réception de **sinistres**
- Le logiciel **n'est pas unique** pour tous les clients, du fait que pas tous les clients ont les mêmes besoins.
- Le processus de construire une solution pour un client est **un build**



The screenshot shows the SHIFT application interface with a table of claims. The table has columns for Status, Alert Number, Claim Number, Policy Number, Score, Suspicion, Claim Cause, Alert Date, Loss Date, Notification Date, and Line of Business. The table contains 10 rows of data, each representing a claim. The first row is highlighted in green, indicating it is the selected claim.

Status	Alert Number	Claim Number	Policy Number	Score	Suspicion	Claim Cause	Alert Date	Loss Date	Notification Date	Line of Business
Unread	DEAG0321248	DEAG0321248	DEPM0212483	10	Manually alerted by user	Accident with identifiable third party	02/07/2024	05/07/2024		Motor
Unread	DEAG0321249	DEAG0321249	DEPM0212483	10	Inconsistent injuries	Accident with identifiable third party	01/07/2024	11/04/2024	28/04/2024	Motor
Unread	DEAG0321250	DEAG0321250	DEPM0212483	10	Inconsistent injuries	Accident with identifiable third party	01/07/2024	11/04/2024	28/04/2024	Motor
Unread	DEAG0321251	DEAG0321251	DEPM0212483	10	Staged accident	Accident with identifiable third party	01/07/2024	05/12/2023	02/12/2023	Motor
Unread	DEAG0321252	DEAG0321252	DEPM0212483	10	ECLA - Staged total loss	Theft of entire vehicle	01/07/2024	24/05/2024	24/05/2024	Motor
Unread	DEAG0321253	DEAG0321253	DEPM0212483	10	ECLA - Staged total loss	Theft of entire vehicle	01/07/2024	24/05/2024	24/05/2024	Motor
Unread	DEAG0321254	DEAG0321254	DEPM0212483	10	ECLA - Staged total loss of vehicle	Theft of entire vehicle	01/07/2024	14/04/2022	15/04/2022	Motor
Unread	DEAG0321255	DEAG0321255	DEPM0212483	10	Staged total loss of vehicle	Theft of entire vehicle	01/07/2024	05/10/2023	04/10/2023	Motor
Unread	DEAG0321256	DEAG0321256	DEPM0212483	10	Staged total loss of vehicle	Theft of entire vehicle	01/07/2024	05/10/2023		Motor
Unread	DEAG0321257	DEAG0321257	DEPM0212483	10	Invalid/Liability/Cover	Accident on collision	01/07/2024	18/05/2024	18/05/2024	Motor

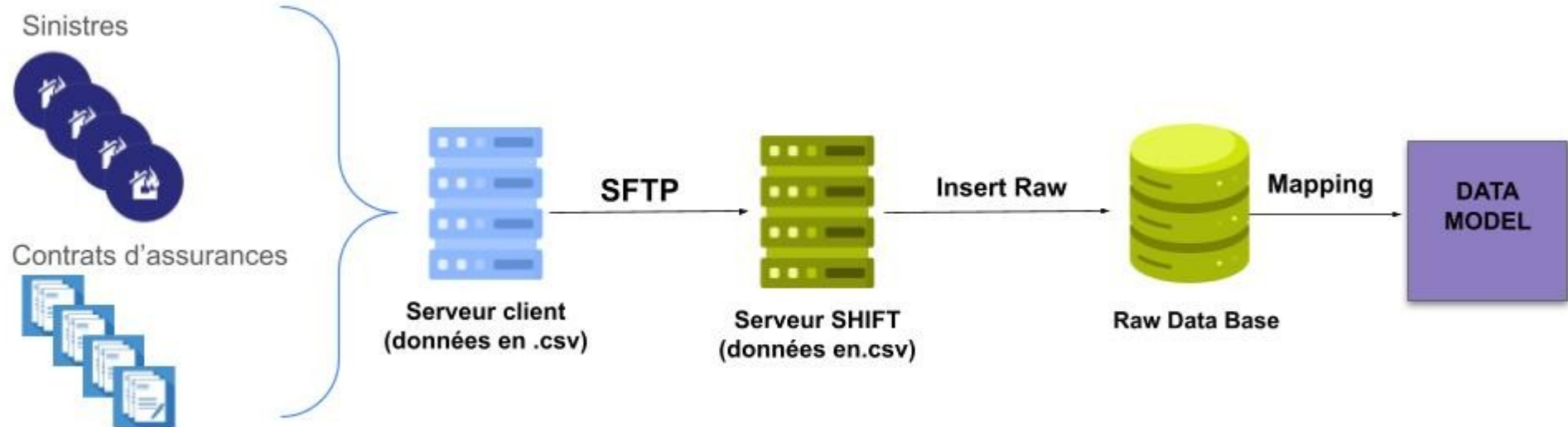


The screenshot shows the SHIFT application interface with a detailed view of a claim. The interface is divided into several sections: Overview, Notes, and Audit Trail. The main section is titled 'SUMMARY' and contains fields for Claim, Policy, Vehicle, and First Party. The Claim section includes fields for Claim Number, Line of Business, Claim Cause, Loss Date, Notification Date, Loss Address, Insured Vehicle, Paid Amount, Reserve Amount, Claim Amount, Status, and Claim Description. The Policy section includes fields for Policy Number, Creation Date, Registration Date, Formula, Policyholder Name, Policy State, and Annual Premium. The Vehicle section includes fields for Make, Model, VIN, First Registration Date, and Party Role. The First Party section includes fields for Name, Address, Bank Account, Date of Birth, Email, and Phone Number. A map of Bari, Italy, is shown on the right side of the interface.

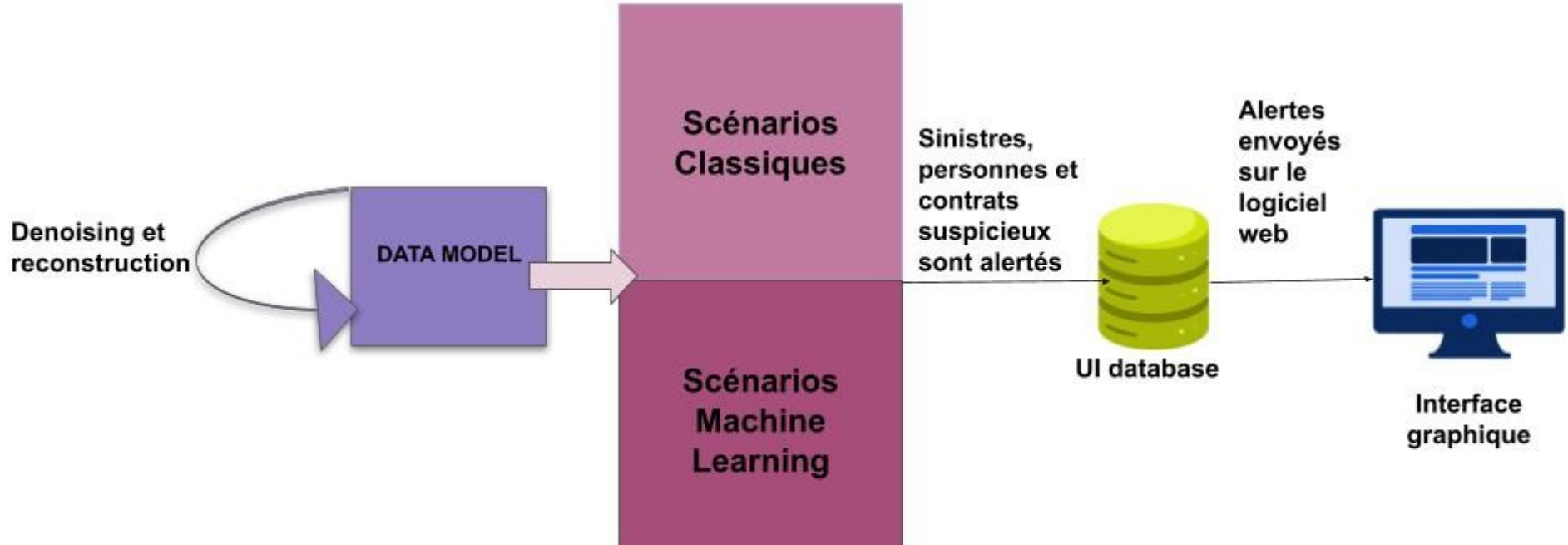
Section	Field	Value
Claim	Claim Number	ITF800000000
	Line of Business	Motor
	Claim Cause	Accident with identifiable third party
	Loss Date	11/04/2024
	Notification Date	28/04/2024
	Loss Address	Corso Trieste, 70124 Bari BA, Italy
	Insured Vehicle	DE 314 84
	Paid Amount	0
	Reserve Amount	40
	Claim Amount	434,379
Policy	Policy Number	ITF800000000
	Creation Date	14/07/2023
	Registration Date	21/07/2023
	Formula	BC Auto
	Policyholder Name	Valentina Mestri
Vehicle	Make	FIAT
	Model	Punto
	VIN	DE 314 84
	First Registration Date	20/02/2012
First Party	Name	Valentina Mestri
	Address	Via Angelo Rinaldi, 1 47, 70124 Bari BA, Italy
	Bank Account	Unknown information
	Date of Birth	12/02/1992
	Email	val_mestri@yahoo.com



# De la réception des données aux alertes I



# De la réception des données aux alertes II



# Latam PnC – Calibration des scénarios

- Les scénarios sont les algorithmes de détection de fraude, ils suivent des règles conditionnelles

ClaimId	Nom	Date du sinistre	Date de souscription	Temps entre la souscription et le sinistre	Le véhicule est de haute gamme	Un PV a été établie	Le sinistre est une fraud avérée	Alerte envoyé
528491	Delon, Alain	01-08-25	30-07-25	2 jours	1	0	1	1
9344	Belmondo, Jean-Paul	01-08-25	05-07-25	26 jours	1	1	0	1
280825	Bardot, Briggite	02-08-25	15-07-25	17 jours	0	0	0	0

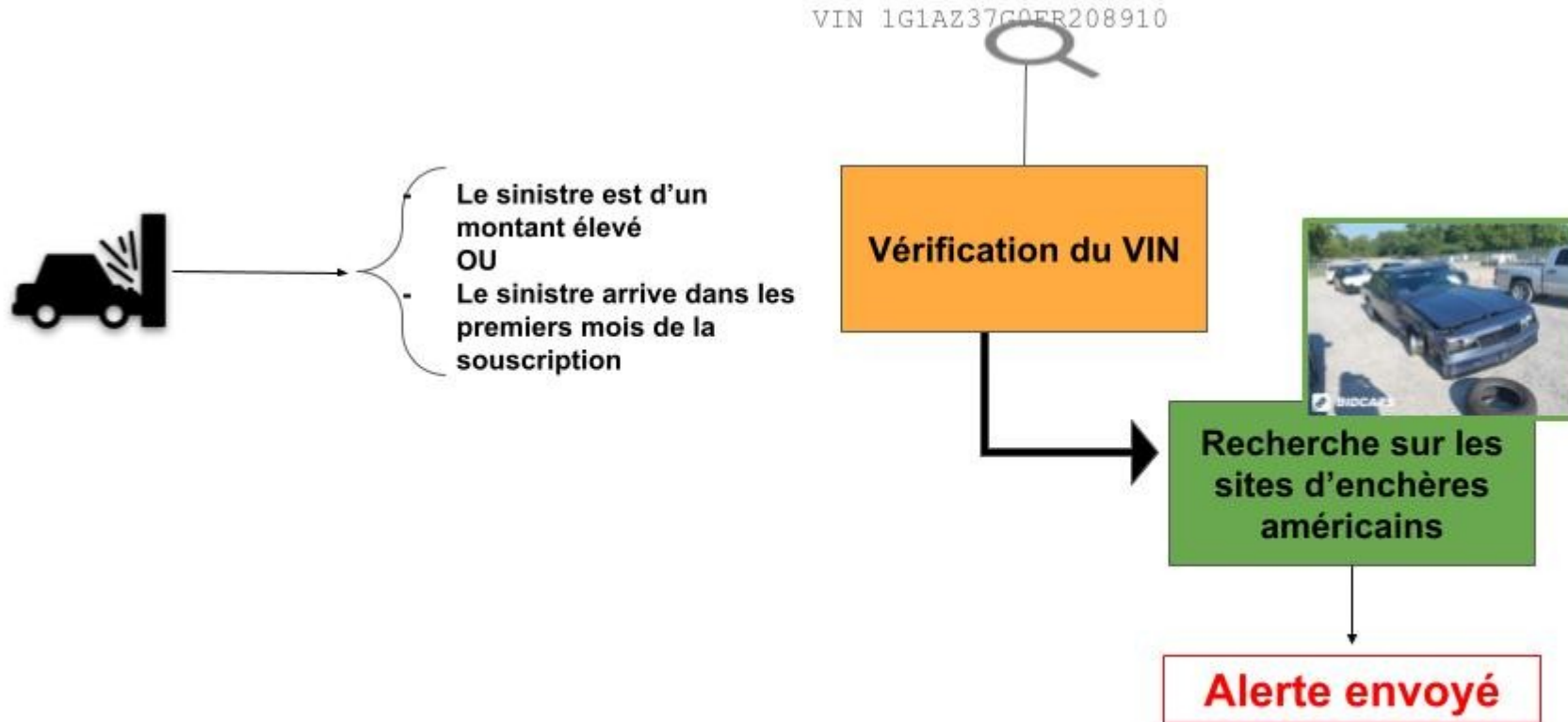
Une alerte est envoyé si toutes les variables primaires sont “activées”, dans le cas des variables secondaires, elles servent à donner une explication à des faits qui pourrait être perçu comme suspicieux

L’objectif des calibration est d’augmenter le nombre d’alertes sans augmenter le nombre de faux positifs

# Latam PnC – Calibration des scénarios

- Au cours du stage nous avons réalisés nombreuses calibration de scénarios; celles-ci sont demandés par le client avant du sprint **jira**
- Une fois la calibration effectué, on présente les test faits en **préproduction**
- Si le client valide, le changement passe à production
- **Exemples de calibration effectués:**
  - Changements dans les scénarios de souscription récente
  - Changement dans le scénario de responsabilité civile
  - Que certains contrats d'assurances ne déclenchent aucune alerte
  - L'envoi automatique des alertes par mails

# Latam PnC - Création de nouveaux scénarios



# Machine Learning

**Objectif** : Entraîner et déployer un modèle de classification binaire pour détecter des fraudes sans motif évident, que les scénarios à règles ne peuvent pas attraper.

**Défis majeurs de la détection de fraude en ML :**

**1) Fort déséquilibre des classes** : Le nombre de sinistres frauduleux (classe minoritaire) est extrêmement faible par rapport aux sinistres légitimes.

**2) Incertitude des labels** : De nombreux sinistres étiquetés "sans fraude" sont en réalité des fraudes non détectées, ce qui "pollue" les données d'entraînement.

Framework choisi : **LightGBM**

# Machine Learning – Génération du dataset

- Nous disposons de deux data set pour l'entraînement: *bodily\_injuries\_fraud* et *material\_damages\_fraud*
- Les données viennent du **data model**, pas de données raw du client

## ***Material Damages Data***

- 131 000 sinistres
- 3.5% des sinistres sont frauduleux

Les données **historiques** contiennent tout 2024

Les données **récentes** contiennent les données de mars, avril et mai 2025

## ***Bodily Injuries Data***

- 21 00 sinistres
- 2.5% des sinistres sont frauduleux

Plusieurs features ne sont pas présentes dans le dataset ce qui a rendu l'entraînement problématique

# ML – Nettoyage des données

**Filtrage des sinistres dupliqués** : suppression des entrées présentant le même identifiant ou des informations identiques sur plusieurs variables clés.

**Uniformisation des types de données** : vérification que les champs numériques, textuels et de dates sont correctement typés dans l'ensemble des fichiers.

**Harmonisation des formats** : par exemple, s'assurer que les dates sont dans un format unique et que les noms de marques ne présentent pas de variations orthographiques

**Vérification des valeurs aberrantes** : détection et traitement des âges, durées ou montants manifestement incohérents.

**Gestion des valeurs manquantes** : identification des champs incomplets et choix d'une stratégie adaptée (imputation, suppression, etc.)

On utilise l'outil de *pandas-profiling* qui peut nous donner des informations précieuses sur le dataset

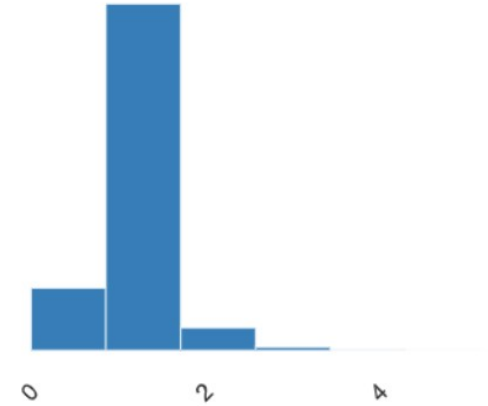


## NumberOfPreviousInvestigatedClaimsWithSimilarCircumstances

Real number ( $\mathbb{R}$ )

Distinct	7
Distinct (%)	0.1%
Missing	598824
Missing (%)	98.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.9266633483

Minimum	0
Maximum	6
Zeros	1723
Zeros (%)	0.3%
Negative	0
Negative (%)	0.0%
Memory size	9.3 MiB



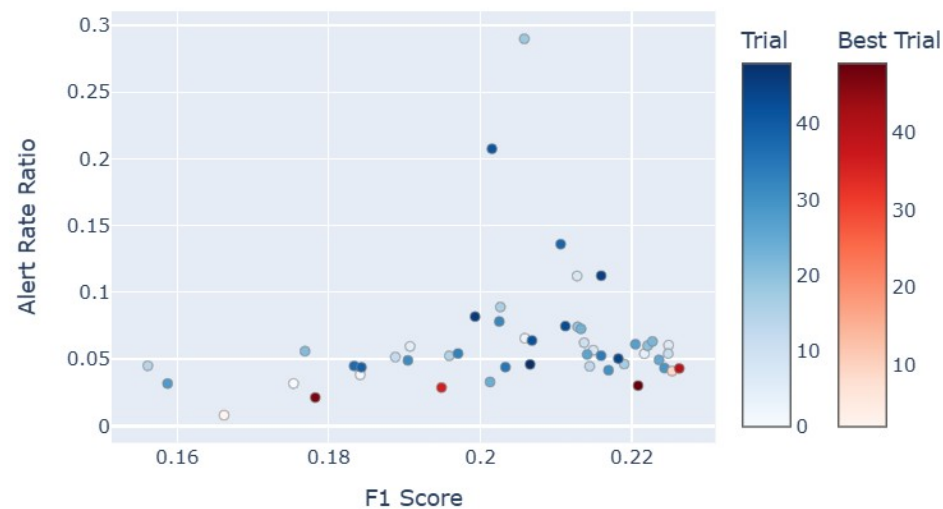
Impossible de regarder toutes les features, ainsi nous prenons un modèle d'un autre client et nous regardons le profiling des 20 premières variables

Après l'entraînement on va re-regarder le profiling pour être certain que les top features n'ont pas beaucoup de valeurs manquantes

# ML – Entraînement

Parameter	Value
metric	F1_score
objective	Binary
verbosity	-1
eta	0.1
feature_fraction	0.9
min_data_in_leaf	10
max_delta_step	1
max_depth	10
lambda_l2	5

Pareto-front Plot

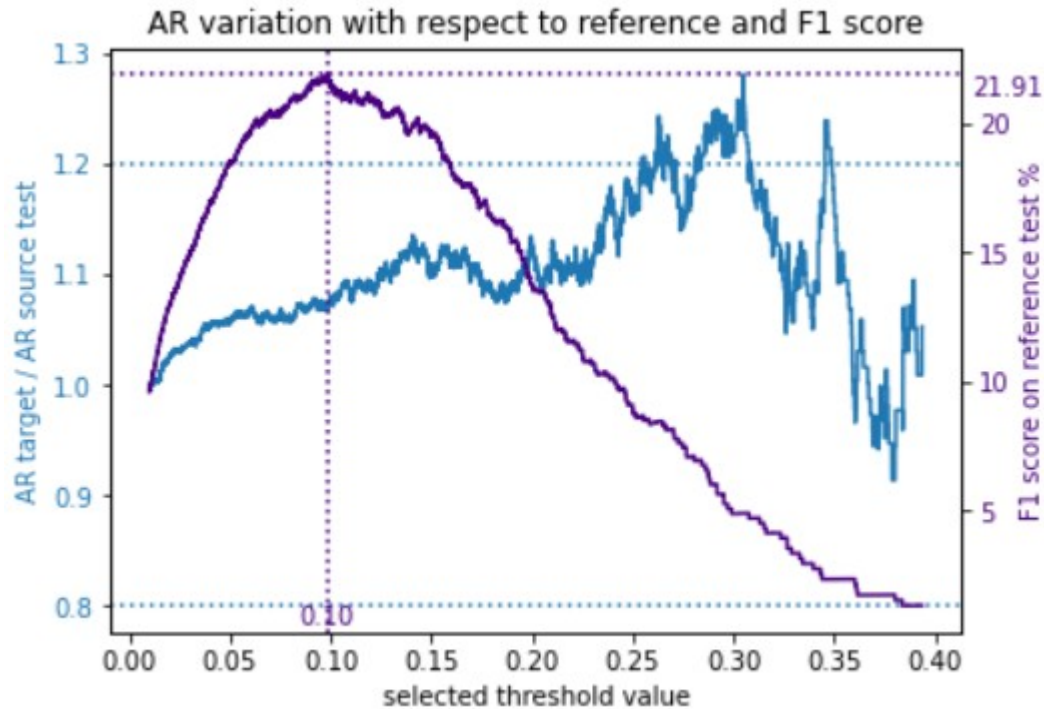


# ML– Choix du modèle

	TN	FN	TP	FP	PP	Pr	Rcl	F1	Alert Rate
Threshold									
0.01	12117	134	922	25567	26489	3.48	87.31	6.69	68.38
0.02	20931	274	782	16753	17535	4.46	74.05	8.41	45.26
0.03	25404	416	640	12280	12920	4.95	60.61	9.16	33.35
0.04	28324	505	551	9360	9911	5.56	52.18	10.05	25.58
0.05	30285	569	487	7399	7886	6.18	46.12	10.89	20.36
0.10	34577	780	276	3107	3383	8.16	26.14	12.44	8.73
0.15	35920	853	203	1764	1967	10.32	19.22	13.43	5.08
0.20	36481	900	156	1203	1359	11.48	14.77	12.92	3.51
0.25	36820	940	116	864	980	11.84	10.98	11.39	2.53
0.30	37078	960	96	606	702	13.68	9.09	10.92	1.81
0.35	37218	980	76	466	542	14.02	7.20	9.51	1.40
0.40	37345	998	58	339	397	14.61	5.49	7.98	1.02
0.45	37440	1012	44	244	288	15.28	4.17	6.55	0.74
0.50	37495	1017	39	189	228	17.11	3.69	6.07	0.59
0.55	37543	1026	30	141	171	17.54	2.84	4.89	0.44
0.60	37586	1032	24	98	122	19.67	2.27	4.07	0.31
0.65	37611	1038	18	73	91	19.78	1.70	3.14	0.23
0.70	37636	1042	14	48	62	22.58	1.33	2.50	0.16
0.75	37653	1046	10	31	41	24.39	0.95	1.82	0.11
0.80	37660	1047	9	24	33	27.27	0.85	1.65	0.09
0.85	37666	1050	6	18	24	25.00	0.57	1.11	0.06
0.90	37675	1051	5	9	14	35.71	0.47	0.93	0.04

	TN	FN	TP	FP	PP	Pr	Rcl	F1	Alert Rate
Threshold									
0.01	6162	96	815	12875	13690	5.95	89.46	11.16	68.63
0.02	10831	227	684	8206	8890	7.69	75.08	13.96	44.57
0.03	13079	337	574	5958	6532	8.79	63.01	15.42	32.75
0.04	14525	417	494	4512	5006	9.87	54.23	16.70	25.10
0.05	15521	494	417	3516	3933	10.60	45.77	17.22	19.72
0.10	17626	681	230	1411	1641	14.02	25.25	18.03	8.23
0.15	18215	754	157	822	979	16.04	17.23	16.61	4.91
0.20	18503	796	115	534	649	17.69	12.62	14.73	3.26
0.25	18664	819	92	373	465	19.78	10.10	13.37	2.33
0.30	18755	842	69	282	351	19.66	7.57	10.94	1.76
0.35	18829	858	53	208	261	20.31	5.82	9.04	1.31
0.40	18892	872	39	145	184	21.20	4.28	7.12	0.92
0.45	18926	879	32	111	143	22.38	3.51	6.07	0.72
0.50	18951	887	24	86	110	21.82	2.63	4.70	0.55
0.55	18961	889	22	76	98	22.22	2.41	4.36	0.50
0.60	18975	895	16	62	78	20.51	1.76	3.24	0.39
0.65	18990	901	10	47	57	17.54	1.10	2.07	0.29
0.70	19002	905	6	35	41	14.63	0.66	1.26	0.21
0.75	19014	906	5	23	28	17.86	0.55	1.06	0.14
0.80	19022	908	3	15	18	16.67	0.33	0.65	0.09
0.85	19031	910	1	6	7	14.29	0.11	0.22	0.04
0.90	19034	911	0	3	3	0.00	0.00	0.00	0.02

# ML-Métriques



$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$ARR = \frac{AlertRate_{recent}}{AlertRate_{historique}}$$

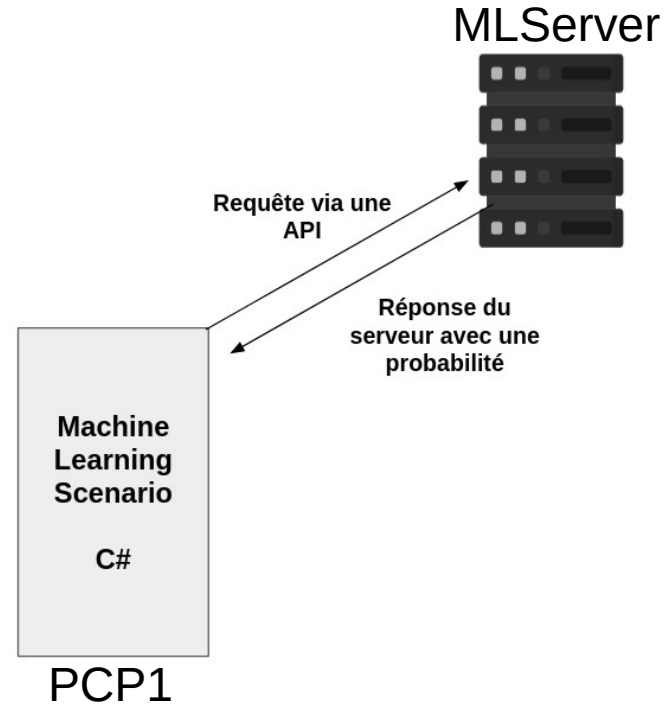
# ML - Mise en pré-production

Le modèle est déployé dans un serveur indépendant de préproduction

Il est composé de plusieurs **JSON** et un **YML**.  
Les .json contiennent les poids du modèle, les features et leur importance et l'ordre dans l'arbre.

Le fichier .yml explicite la localisation des fichiers .json

Une API propre de Shift gère l'interaction entre les deux outils.



# ML - Mise en production

- Le scénario est testé en préproduction pour être certain que des alertes sont bien envoyés
- Un threshold est choisi en fonction des alertes qui sont vraiment envoyés
- On regarde les premières 20 sinistres, et on regarde les indices de Shapley pour être certain qu'en pratique une seule variable ne "monopolise" pas le modèle
- J'ai regardé toutes les alertes du modèle des dommages corporels et une cinquantaine du modèle de dommages matériels
- Une fois les alertes pertinentes triées, on présente les alertes et la volumétrie au client
- Le modèle est déployé à production et le code mergé à **develop**

# Cas de démonstration

**SHIFT** Alerts Exploration Reporting Smart Documents

Q Search

Score: 83 Alert reference: ITFB1010216 Scenario: Inconsistent injuries Alert Date: 28/04/2024

Overview Notes Audit Trail

**SUMMARY** Show less...

**Claim**

Claim Number	ITFB1010216
Line of Business	Motor
Claim Cause	Accident with identified third party
Loss Date	13/04/2024
Notification Date	28/04/2024
Loss Address	Corso Trieste, 70126 Bari BA, Italy
Insured Vehicle	SV-354-BM
Paid Amount	€0
Reserve Amount	€0
Claim Amount	€164,339
Status	Open
Claim Description	Eravamo in coda e il veicolo davanti s...

**Policy**

Policy Number	P74635928
Creation Date	14/07/2021
Inception Date	21/07/2021
Expiration Date	01/01/2025
Formula	RC Auto
Policyholder Name	Valentina Moretti
Policy State	Unknown information
Annual Premium	Unknown information

**Vehicle**

Make	Fiat
Model	Punto
VRN	SV-354-BM
First Registration Date	20/12/2012
VIN	Unknown information
Party Side	First Party

**First Party**

Name	Valentina Moretti
Address	Via Angelo Bassi, 1-17, 70124 Bari BA,...
Bank Account	Unknown information
Date of Birth	12/03/1982
Email	va_l_moretti@yahoo.com
Phone Number	314 9092185

**Third Party**

# Recul d'expérience

- Pendant ce stage j'ai pu apprendre à:
  - Travailler dans un code complexe et robuste, dans lequel des centaines de personnes contribuent.
  - A écrire du code propre, lisible et des PR structurés
  - A comprendre toutes les étapes d'un modèle de ML, et comment celui-ci est mis en production
  - Comprendre le fonctionnement d'une entreprise, les différentes équipes et leurs interactions
  - Avoir des bases très solides en C#, POO, SQL et Git
  - Surtout, avoir de la responsabilité vers le travail et un engagement avec les dates impartis