Projet 7 :
Développez une preuve de concept : Détection d'anomalies en assurances

OPENCLASSROOMS

Cédric Randrianarivélo

I. Contexte

II. Méthodologie

III. Traitement des données

IV. Modélisation

V. Présentation des résultats

OPENCLASSROOMS

I.Contexte

OPENCLASSROOMS

Contexte

J'ai effectué mon parcours « Machine Learning Engineer » chez Openclassrooms, avec une Mission d'entreprise chez l'entreprise Grope BNP PARIBAS dans l'entité cardiff dans le secteur de l'assurances.

Dans le département production financière au sein de l'équipe « Data Analyse & Innovation ». Avec pour objectif d'aider les équipes métier comme l'actuariat, contrôle de gestion, comptabilité et équipe RISK.

L'idée était d'aider les différentes équipes à digitaliser les processus « Manuel ». En développement des outils de traitement de la donnée et de calculs.

Ensuite, nous avons procédé au déploiement des outils et avons assuré la montée en compétences des différentes équipes métier quant à leur utilisation. De plus, nous avons pris en charge la maintenance de ces outils.

OPENCLASSROOMS

Problématique

Il y a eu le constat qu'un grand nombre de données reçu possédaient des anomalies et incohérences lors de son chargement. (ex : Montant élevé, données manquantes…)

Une partie de la mission consistait à mettre en place un système de contrôle de qualité afin de détecter les lignes de données anormales.

La décision de partir sur un système de machine pour détecter les anomalies a été prise. Pour anticiper les anomalies futures.

Nous souhaitions faciliter la tâche des équipes métier et améliorer l'efficacité des processus et de leurs qualités.

Objectifs

Objectifs de la mission d'entreprise :

D'AUGMENTER LA QUALITÉ ET LA FIABILITÉ DES DONNÉES

DE RENDRE LA TÂCHE DE VÉRIFICATION DES DONNÉES AUTOMATIQUE ET MOINS CHRONOPHAGE

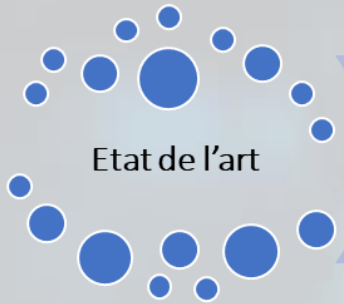
LIVRER UN OUTIL DE DÉTECTION D'ANOMALIES PAR MACHINE LEARNING.

FAIRE MONTER EN COMPÉTENCE LES ÉQUIPES MÉTIERS SUR L'OUTIL

II. Méthodologie

OPENCLASSROOMS

Plan chronologique



Etat de l'art

Mise en perspective des études faites sur la détection d'anomalie en assurances

Data preparation

Traitement et homogénéisation des données

Feature engineering

Variables supplémentaires créées à partir de données initiales permettant d'identifier des anomalies

Annotation des données

Déterminer les lignes de données considérées comme anomales

Modélisation

Développement des algorithmes d'apprentissage

Mise en production

Création d'une interface de détection d'anomalies

Etudes bibliographique

Modèle de base: Dummy Classifier

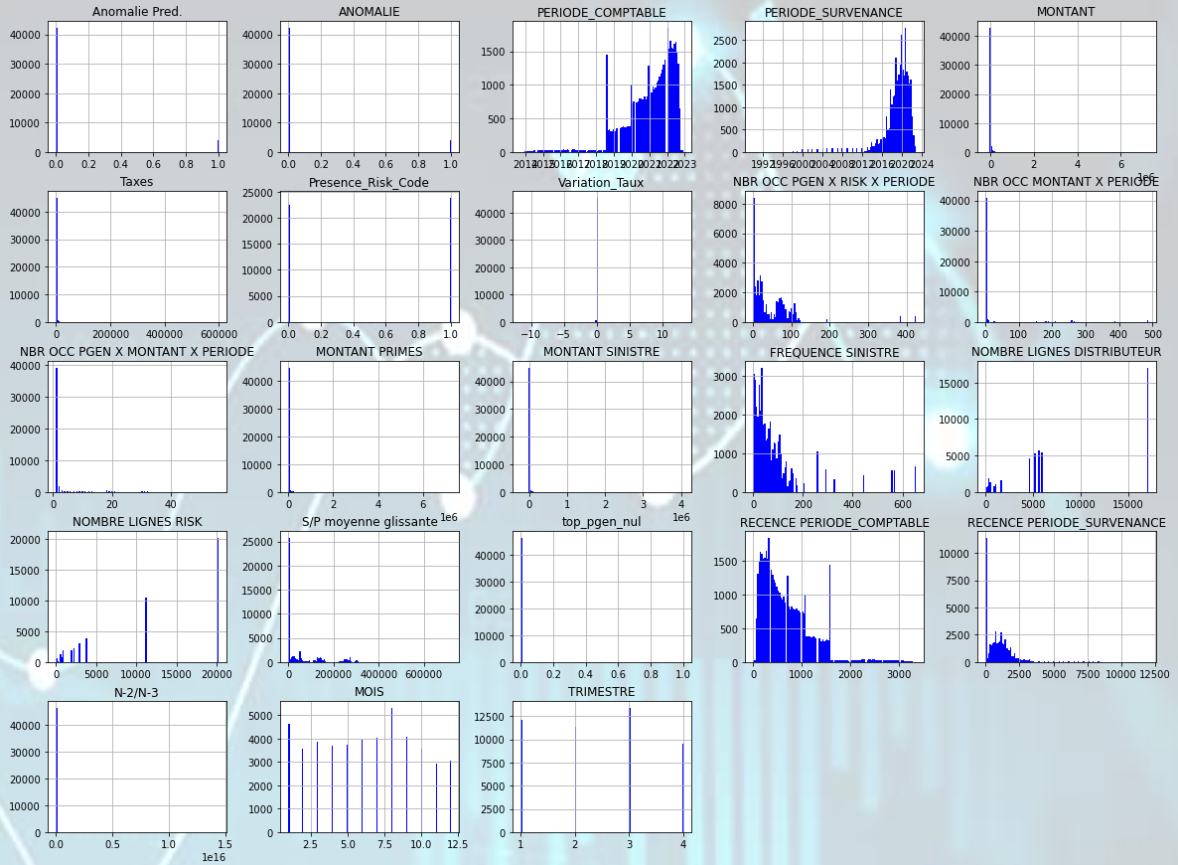
Modéles upervisés	Modéles non-supervisés
Decision trees	K-means
Random forest	Isolation forest
Gradient Boosting	Self-organizing maps (SOM)
Modèles linéaires	C-means
Multi-layer perceptron	Adaptive resonance theory (ART)
Support vector machine Learning	Local outlier factor (LOF)

III. Traitement des données

OPENCLASSROOMS

Données

- **Base de données** contenant des informations sur les sinistres et les primes. (primes : versements mensuels des assurés et sinistres : montants versés par les assurances quand il y a un litige)
- Base de données **non homogène**. Plusieurs sources non homogènes, provenant de différents délégataires externes.
- **Travail d'homogénéisation** : Concaténer et homogénéiser ces données pour les rendre cohérentes et uniformes.



Feature engineering

Features
La variation : N/N-1
La saisonalité : (N-2)-(N-1) /2 (N-3)-(N-1) /2 sur 12 derniers mois par risque (la moyenne de deux mois sur trois mois glissants)
La saisonalité : (N-2)-(N-1) /2 (N-3)-(N-1) /2 sur 12 derniers mois par distributeur (la moyenne de deux mois sur trois mois glissants)
Prime périodique/unique
Indicateur Nouveau PGEN 0/1
Nombre d'occurrence d'un même PGEN x RISK sur une période donnée
Nombre d'occurrence d'un même montant sur une période donnée
Nombre d'occurrence d'un même montant sur une période donnée et un PGEN donné
PGEN vide
Risque présent ou absent du référentiel APLE
Fréquence de sinistres
Fréquence de sinistre remboursés

Feature engineering

DATA EXPLORATION :	**ANALYSE DES DISTRIBUTIONS DES VARIABLES** **DÉTERMINATION DES VARIABLES PERTINENTES (AVIS MÉTIERS ET DES EXPERT)**
ANALYSE EN COMPOSANTES PRINCIPALES (ACP) :	**DÉTERMINER LE NOMBRE DE FEATURES SIGNICATIVES DANS NOTRE JEU DE DONNÉES** **85% DE L'EXPLICATION DU JEU DE DONNÉES AVEC 16 FEATURES**
MATRICE DE CORRÉLATION :	**DÉTERMINATION DES FEATURES CORRÉLÉES** **ELIMINATION DE LA REDONDANCE ET RÉDUIRE LE NOMBRE DE FEATURES.**

À la fin de cette étape, nous avons réussi à réduire le nombre de features de 43 à 16.

IV. Modélisation

OPENCLASSROOMS

Création de la variable cible

Annotation des données :

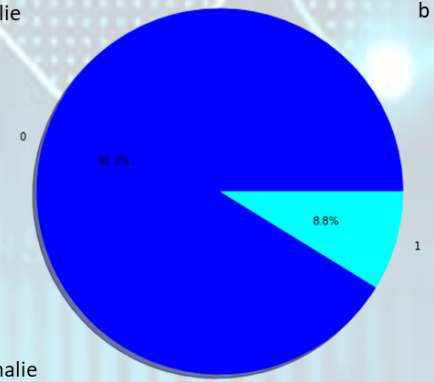
Avec Utilisation de Power BI pour identifier les points considérés comme des anomalies en termes de points. En passant par des méthodes d'identification statistiques.

Règle de gestion :

Mise en évidence des cas où les PGEN (paramètres généraux) sont manquants. Mise en évidence du RISK manquant dans la ligne.

A : Montant anomalie

b : Montant sans anomalie



C : Pourcentage Anomalie

Modèlisation

Mise en place des algorithmes supervisés et non supervisés sur les jeux de données

Un modèle Non supervisé (K-means et Isolation Forest) non performant. (Isolation des gros montants…)

Décision de se concentrer sur les modèles supervisés

Modèle SVM et Linéaire non prometteurs

Modèlisation

Modèle MLP est intéressant mais son exécution et l'optimisation des paramètres est trop long

Modèle de Gradient Boosting et Random Forest Très pertinents

Nous souhaitions optimiser le recall pour capter le plus d'anomalies possible

Modèle Random Forest le plus pertinent dans ce cas là

Déterminer un seuil de probabilité de classes prédites adapté

V. Présentation des résultats

OPENCLASSROOMS

Modèle retenu

	Modèle	Résultat	Validé
Modèle Supervisé	Ridge	Relation trop complexe	X
	Support Vector Machine	Relation trop complexe	X
	Multi-Layer Perceptron	Temps d'execution et d'optimisation trop long	X
	Gradient Boosting	Très bon résultat	X
	Random Forest	Meilleur résultat	O
Modèle non-supervisé	K-Means	Regroupement des petits et grands montants	X
	Isolation Forest	Isolation des sinistres et des primes	X

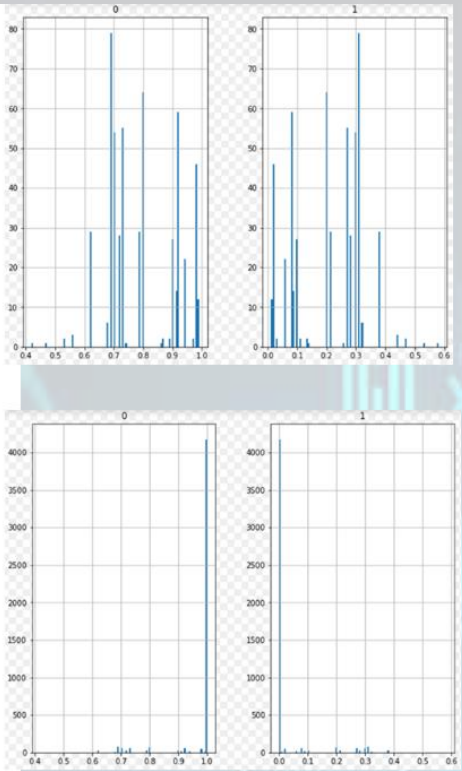
Modèle de Random Forest retenu

	fit_time	score_time	test_accuracy	test_f1	test_precision	test_recall	train_accuracy	train_f1	train_precision	train_recall	Model
0	11.766948	0.029199	0.867945	0.553917	0.572834	0.786416	0.996402	0.976283	0.989049	0.963951	XG Boost
0	0.217104	0.032600	0.922808	0.000000	0.000000	0.000000	0.922808	0.000000	0.000000	0.000000	Ridge
0	42.412970	0.092601	0.920481	0.001172	0.202439	0.000622	0.923197	0.010146	0.994872	0.005127	MLP
0	8.633304	0.061000	0.894572	0.707224	0.651495	0.906468	1.000000	1.000000	1.000000	1.000000	Random Forest

Changement de seuil

	0	1
275	0.98977	0.01023
276	0.98977	0.01023
277	0.89977	0.10023
278	0.98977	0.01023
297	0.98977	0.01023
...	...	...
4539	0.69000	0.31000
4540	0.79000	0.21000
4541	0.68000	0.32000
4542	0.68000	0.32000
4543	0.68000	0.32000

539 rows × 2 columns



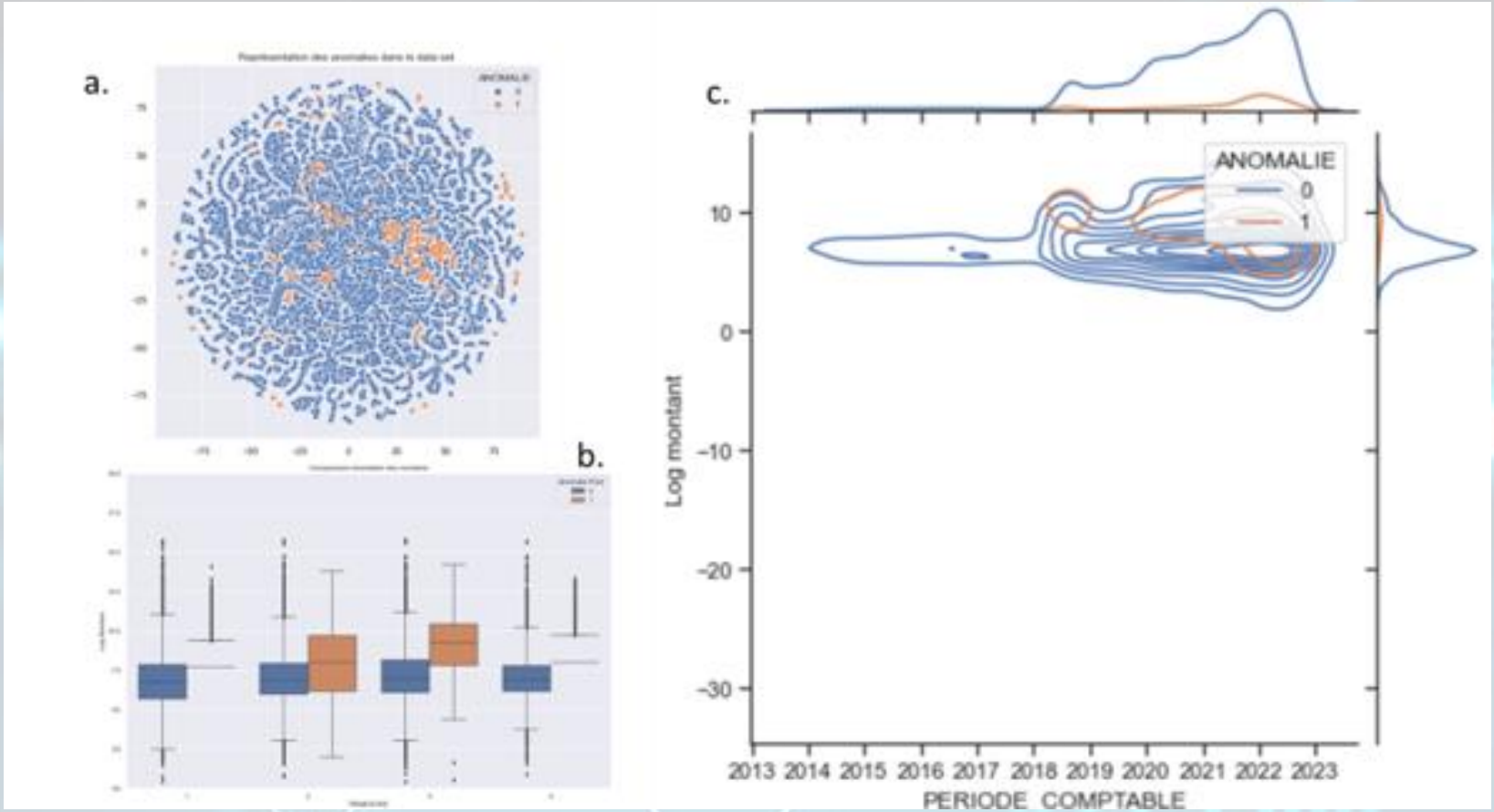
- Une approche sur les **probabilités des classes obtenues**
- Nous avions récupéré les **probabilités prédites**
- Pour gagner en précision sur modèle on a tenté **plusieurs seuils** de probabilité
- Des seuils de **0,6 à 0,3** sur la classe 1
- Pas de gros impacts sur les **seuils 0,6 à 0,4**
- Voir avec le métier le seuil **maximum tolérable**

	test_precision	test_recall	Nb_0	Nb_1	Seuil
0	0.662345	0.894236	4142	393	0.6
1	0.651495	0.906468	4136	399	0.5
2	0.648470	0.917468	4122	413	0.4
3	0.716035	0.821318	3771	582	0.3

Interface

ANOMALIE PREDITE
0
0
1
0
0
0
1
0
0
0
1
0

Anomalies déterminées



Les axes d'améliorations

Feature Engineering	Développer d'autres variables pertinentes pour notre modèle. Améliorer les features présentes.
Approche Non Supervise	Mettre l'accent sur les modèles supervisés en les optimisant.
Apprentissage continu des nouvelles anomalies	Récupération des nouvelles anomalies vérifiées. Apprentissage en continu du modèle.
Distinctions sinistres et primes	Mettre en place plusieurs modèles permettant de distinguer les sinistres et le primes

Questions ?

OPENCLASSROOMS

Merci pour votre attention

OPENCLASSROOMS