
Description sommaire des données

Sujet / Jeu de données 1 : Détection de clients sensibles à la fraude via phishing

Scénario de fraude : Le fraudeur se fait passer pour le Crédit Agricole pour extirper des informations de connexion aux applications du Crédit Agricole. Une fois connecté sur le compte du client, le fraudeur peut effectuer des opérations via Carte ou Virement.

L'idée est de détecter ici les clients sensibles à ce type de fraude.

Volumes / périmètre :

	Toulouse 31	Languedoc
Nombre clients total	409 000	910 000
Nombre clients fraudés	523	1 200

Variables explicatives :

- Données signalétiques : Âge, ancienneté, CSP, statut marital, code postal...
- Équipement bancaire
- Habitudes sur les outils digitaux
 - Environ 300 indicateurs au total

Le flux bancaire n'est pas visible

L'évènement à prédire est rare, des méthodes de redressement, sur-échantillonnage, etc. sont à prévoir sur les données.

Les méthodes habituelles de prédiction d'un évènement, comme régression logistique, arbres de classification ou SVM ont des difficultés à modéliser des jeux de données déséquilibrés. Pour résoudre ces problèmes, trois types de méthodes sont à notre disposition pour améliorer la qualité prédictive : corrections apportées aux modèles, échantillonnage de données (Undersampling, Oversampling tel que SMOTE) ou méthodes d'agrégation (Bagging, boosting).

Sujet / Jeu de données 2 : Contribution des parcours digitaux sur la souscription d'un crédit consommation

Intérêt : Le crédit conso est un axe stratégique majeur et le digital est un levier fort de développement. Grâce au dispositif de tracking de notre site web ainsi qu'aux outils et méthodes du « big data », le Crédit Agricole cherche à éclairer des zones d'ombre :

- Contribution du digital à la production
- Identification des leviers du parcours client générant de la production

Volumes / périmètre :

	Toulouse 31	Languedoc
Nombre clients ayant souscrit un crédit conso sur les 6 derniers mois	6 700	15 000
Nombre de pages consultées sur 45j (clients reconnus)	6 706 k	14 922 k

Variables explicatives :

- La production de crédits conso sur les 6 derniers mois
- Les données signalétiques : Âge, ancienneté, CSP, statut marital, code postal, score conso...
- Les données de webtracking sur 45 jours
- Les opportunités de contacts (Emails, SMS, Lead...)

→ Environ 300 indicateurs au total

Les données sont à date de la demande de crédit, avec les 45 jours précédents de webtracking.

Groupe 1

Détection d'anomalie supervisée

[Sujet : détection de clients sensibles à la fraude]

Tuteur enseignant : Philippe Berthet (philippe.berthet@math.univ-toulouse.fr)

Chef de groupe : Pauline Dupuy

Sous-Chef / resp. visualisation : Jules Boutibou

Description du groupe

Quels sont les facteurs de risque de fraude ?

- Statistiques descriptives (résultats pouvant être partagés)
- Feature engineering (qui permettent d'améliorer l'explicabilité)
- Explicabilité (on veut des modèles simples mais très explicables)
- Visualisation

Résultats à comparer avec ceux du groupe 2

Données en entrée

- Jeu de données n°1

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Tentative de description des personnes fraudées.
- Réduction de dimension : pourquoi ici, et comment ?
- Interprétation et représentation des regroupements de variables.
- État de l'art sur quelques méthodes supervisées.
- Évaluation de ces méthodes en situation d'événement rare.

Groupe 2

Détection d'anomalie non supervisée

[Sujet : détection de clients sensibles à la fraude]

Tuteur enseignant : Edouard Pauwels (edouard.pauwels@irit.fr)

Chef de groupe : Théo Vedis

Sous-Chef / resp. visualisation : Nesrine Aider

Description du groupe

Est-ce que les anomalies correspondent à des fraudes ?

- Statistiques descriptives (résultats pouvant être partagés)
- Mise en oeuvre et comparaison d'algorithmes de détection d'anomalies
- Feature engineering
- Explicabilité
- Visualisation

Données en entrée

- Jeu de données n°1 (sans utiliser l'indicateur de fraude pour l'apprentissage)

Livrables attendus

- Documentations sur les méthodes mises en oeuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Liste préliminaire de méthodes à évaluer / mettre en oeuvre dans le cadre de ce projet.
- Comment caractériser les sorties de ces algorithmes? Quelles sont les grandes tendances?
- Comment évaluer les performances des méthodes en termes de détection de sensibilité à la fraude?
- Proposer des pistes d'améliorations dans l'objectif de détection de sensibilité à la fraude.

Groupe 3

Clustering

[Sujet : détection de clients sensibles à la fraude]

Tuteur enseignant : Mathieu Serrurier (mathieu.serrurier@irit.fr)

Chef de groupe : Théo Saccareau

Sous-Chef / resp. visualisation : Thierno Diallo

Description du groupe

Constituer des profils de clients

Est-ce qu'il y a un groupe qui est plus sensible à la fraude ? *et éventuellement aux achats de crédit ?*

- Statistiques descriptives (résultats pouvant être partagés)
- Feature engineering
- Explicabilité
- Visualisation des résultats

Données en entrée

- Jeu de données n°1

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- État de l'art sur les algorithmes de clustering
- Métriques d'évaluation
- Outils d'interprétabilité des algorithmes

Groupe 4

Intimité différentielle

[Sujet : détection de clients sensibles à la fraude]

Tuteur enseignant : Adrien Mazoyer (adrien.mazoyer@math.univ-toulouse.fr)

Chef de groupe : Mélina Audiger

Sous-Chef / resp. visualisation : Antoine Godin

Description du groupe

Est-il possible de déduire certaines données en ne regardant que certaines autres ?

Quel est le nombre minimal de variables nécessaires pour détecter la fraude ?

- Statistiques descriptives (résultats pouvant être partagés)
- Visualisation des résultats

Données en entrée

- Jeu de données n°1

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Etat de l'art sur l'intimité ou confidentialité différentielle

Groupe 5

Génération de nouvelles observations

[Sujet : détection de clients sensibles à la fraude]

Tuteur enseignant : Mathieu Serrurier (mathieu.serrurier@irit.fr)

Chef de groupe : Damien Sonnevile

Sous-Chef / resp. visualisation : Stevan Stricot

Description du groupe

Construire un générateur de données qui a les mêmes caractéristiques que le jeu de données initial sans les données originelles

Evaluer les algorithmes sur ces données simulées pour évaluer la qualité du générateur

- Statistiques descriptives (résultats pouvant être partagés)
- Sampling sur les variables avec hypothèse d'indépendance
- Sampling gaussian multivarié
- Évaluation :
 - statistiques descriptives
 - avec un algorithme qui essaie de détecter les données générées par rapport aux vraies données
- Adversarial network

Données en entrée

- Jeu de données n°1

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Etat de l'art sur

Groupe 6

Statistiques descriptives et recherche de patterns

[Sujet : contribution des parcours digitaux à la souscription de crédit conso]

Tuteur enseignant : Dominique Bontemps (dominique.bontemps@math.univ-toulouse.fr)

Chef de groupe : Anaïs Andrieu

Sous-Chef / resp. visualisation : Jordi Mora Fernandez

Description du groupe

- Statistiques descriptives :
 - Qu'est-ce qui décrit le parcours d'un client de façon pertinente ? Comment le généraliser sur l'ensemble des données ?
 - Comment lie-t-on les données clients aux données de son parcours ?
- non supervisé :
 - Clustering de parcours
 - Recherche de pattern sur l'historique de navigation (sous-séquences fréquentes et rares)
 - Comparaison entre les résultats obtenus aux points précédents et les données client
- Visualisation des résultats

Données en entrée

- Jeu de données n°2

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire (à rendre sous Moodle avant les vacances de Noël)

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- État de l'art sur le clustering de parcours, et sur la recherche de pattern.
- Chercher et comprendre l'algorithme apriori dans la littérature.

Groupe 7

Représentation des parcours

[Sujet : contribution des parcours digitaux à la souscription de crédit conso]

Tuteur enseignant : Emmanuelle Claeys (Emmanuelle.Claeys@irit.fr)

Chef de groupe : Michael Corbeau

Sous-Chef / resp. visualisation : Vincent Barudio

Description du groupe

Comment représenter le parcours digital des clients à partir de données de webtracking ?

- Visualisation par diagramme de Sankey
- Réduction de dimensions \Rightarrow clustering ?
- Animation temporelle
- Simulation de parcours par Réseaux de Petri
- Process mining
 - Heuristic Miner
 - Alpha Miner

Données en entrée

- Jeu de données n°2

Livrables attendus


- Documentations sur les méthodes mises en oeuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire (à rendre sous Moodle avant les vacances de Noël)

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Installation Package Python <https://pm4py.fit.fraunhofer.de/>
- Consulter la doc <https://pm4py.fit.fraunhofer.de/documentation#discovery> (Python)
- Vidéos à suivre  [pm4py tutorials - tutorial #1: What is Process Mining?](#)
- Installation de Prom <https://www.promtools.org/doku.php> pour Clustering

Groupe 8

Prédiction de souscription (sans données de webtracking)

[Sujet : contribution des parcours digitaux à la souscription de crédit conso]

Tuteur enseignant : Cécile Chouquet (cecile.chouquet@math.univ-toulouse.fr)

Chef de groupe : Vincent Blase

Sous-Chef / resp. visualisation : Nour Elhouda Kired

Description du groupe

Quelles sont les caractéristiques des souscripteurs ?

- Statistiques descriptives (résultats pouvant être partagés)
- Feature engineering (qui permettent d'améliorer l'explicabilité)
- Explicabilité (on veut des modèles simples mais très explicables)
- Visualisation
- Comparaison semi-supervisée avec le jeu de données n°1

Résultats à comparer avec ceux du groupe 9

Données en entrée

- Jeu de données n°2 (sans les données de webtracking)

Livrables attendus

- Documentations sur les méthodes mises en oeuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire (à rendre sous Moodle avant les vacances de Noël)

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Réduction de dimension : pourquoi ici, et comment ?
- Interprétation et représentation des regroupements de variables.
- État de l'art sur quelques méthodes supervisées.
- Évaluation de ces méthodes en situation d'événement rare.

Groupe 9

Prédiction de souscription (avec données de webtracking)

[Sujet : contribution des parcours digitaux à la souscription de crédit conso]

Tuteur enseignant : Emma Grandgirard (emma.grandgirard@univ-tlse3.fr)

Chef de groupe : Karine Biard

Sous-Chef / resp. visualisation : Marianne Manson

Description du groupe

Quel est le parcours type de la souscription ? → visites web, contacts banque...

Quelles sont les variables explicatives de la souscription ?

Les données de webtracking permettent-elles d'améliorer la prédiction de la souscription ?

Estimer la longueur (temps) d'historique webtracking nécessaire pour prédire la souscription

- Statistiques descriptives (résultats pouvant être partagés)
- Feature engineering (permettant d'améliorer l'explicabilité) avec focus sur le webtracking
- Explicabilité (on veut des modèles simples mais très explicables)
- Evaluation en situation d'événement rare
- Visualisation

Résultats à comparer avec ceux du groupe 8

Données en entrée

- Jeu de données n°2

Livrables attendus

- Documentations sur les méthodes mises en œuvre, évaluées et choisies
- Notebooks commentés
- Visualisation et interprétation des résultats dans un document de synthèse

Travail préparatoire (à rendre sous Moodle avant les vacances de Noël)

À rendre sous Moodle avant les vacances de Noël :

- Brown papers et tous les documents fournis pour le travail avec le catalyseur
- Travail préparatoire demandé aux L3, M1 et M2 (tutos à lire, méthodes à comprendre, etc.)

À faire en plus pour monter en compétences :

- Rechercher / installer des librairies d'explicabilité (ex : SHAP, Lime, Alibi)
- Rechercher des méthodes d'évaluation en situation d'événement rare
- Explorer le [Bank Marketing Dataset](#) et les notebooks Kaggle associés (entraînement à la réduction de dimension et à l'utilisation de méthodes supervisées)
- Faire un état de l'art sur l'utilisation de données webtracking en machine learning

Travail préparatoire des M2 non alternants

(4 personnes)

Charte de codage (à reprendre et adapter des années précédentes, en particulier celle de 2019-20)

Travail sur les fichiers de données :

- Organisation et mise en forme du jeu de données n°2
- Dans quelle mesure y a-t-il une intersection entre les 2 jeux de données ? Y a-t-il un intérêt à les concaténer ?

Mise en place d'une plateforme globale (site web par exemple) de présentation et visualisation des résultats. Cette plateforme sera ensuite reprise par chaque groupe pour présenter ses propres résultats, il est donc extrêmement important de bien documenter chacune des parties et bibliothèques utilisées.

Attention : ici, la communication avec les sous-chefs responsables de la visualisation est fondamentale.