27 novembre 2023

Techniques et méthodes de scoring

-

Projet 2023-2024

-

3e année – Gestion des Risques

LCL BANQUE ET ASSURANCE

Ma vie. Ma ville. Ma banque.

Introduction

Informations pratiques liées au Projet de Techniques et Méthodes de Scoring

- Présentation de l'intervenant ;
  - Mail : clement.ducroquetz@lcl.fr

- 3 x 6h de TD :
  - 27/11/23 : de 09h45 à 12h45 et de 14h à 17h ;
  - 04/12/23 : de 09h45 à 12h45 et de 14h à 17h ;
  - 11/12/23 : de 09h45 à 12h45 et de 14h à 17h ;

- Projet à réaliser par groupe de 3 étudiants.

- Rapport à remettre au plus tard fin janvier 2023 (date prévisionnelle à confirmer).

- Présentation des meilleurs rapport devant les métiers d'LCL au T1-2023.

Introduction

Attendus concernant le projet

- Calibrer un **modèle de score à l'aide d'une régression logistique** sur données réelles, en mesurer les performances, puis construire une échelle de rating mesurant le risque intrinsèque des clients.
- Challenger ce modèle par des techniques de Machine Learning.

- Les résultats de ce projet devront être consignés dans un rapport de 20 pages maximum qui **détaillera la démarche empruntée, les résultats obtenus et leurs interprétations**.

- Plan indicatif du rapport :
  - ✓ Introduction ;
  - ✓ Analyses descriptives / Nettoyage base / Transformation de variables ;
  - ✓ Construction des bases / Echantillonnage ;
  - ✓ Sélection des variables / Discrétisation / Regroupement ;
  - ✓ Estimation du modèle ;
  - ✓ Analyse des performances ;
  - ✓ Elaboration de la grille de score ;
  - ✓ Modèles challengers ;
  - ✓ Conclusion / préconisation.

- Un Powerpoint sera réalisé en collaboration entre les groupes retenus afin de synthétiser les différents travaux et de les présenter aux métiers de LCL.

1

Présentation de LCL

2

Présentation des activités de RCP\Modélisation

3

Présentation des grandes étapes de construction d’un score

4

Présentation du projet

5

Ateliers

Présentation de LCL

LCL & le Groupe Crédit Agricole

Ma vie. Ma ville. Ma banque.

Banque de proximité en France

banque & assurances

BANK

Ma vie. Ma ville. Ma banque.

Interfimo

Banque de financement et d'investissement

CRÉDIT AGRICOLE CORPORATE & INVESTMENT BANK

Crédit à la consommation

CRÉDIT AGRICOLE CONSUMER FINANCE

Agos DUCATO

Sofinco BANK

Gestion d'actifs

Amundi ASSET MANAGEMENT

caceis INVESTOR SERVICES

Banque de proximité à l'international

CRÉDIT AGRICOLE CARIPARMA | FRIULADRIA | CARISPEZIA

CRÉDIT AGRICOLE
Pologne · Roumanie · Ukraine
Serbie · Egypte

مصرف المغرب
CRÉDIT DU MAROC
Toute une banque pour vous

Immobilier

CRÉDIT AGRICOLE IMMOBILIER

ANGLE NEUF

SQUARE HABITAT CRÉDIT AGRICOLE

Assurances

CRÉDIT AGRICOLE ASSURANCES

viavita

Banque Privée

LCL BANQUE PRIVÉE

CRÉDIT AGRICOLE PRIVATE BANKING

INDOSUEZ WEALTH MANAGEMENT

CFM INDOSUEZ WEALTH MANAGEMENT

Activités spécialisées

CRÉDIT AGRICOLE CAPITAL INVESTISSEMENT & FINANCE

CA GRANDS CRUS

unimédias

pleinchamp.com
Votre site d'expertise agricole.

CRÉDIT AGRICOLE PAYMENT SERVICES

Crédit bail & affacturage

CRÉDIT AGRICOLE LEASING & FACTORING

EUROFACTOR GROUPE CRÉDIT AGRICOLE

UNIFERGIE GROUPE CRÉDIT AGRICOLE

Présentation de LCL

LCL & le Groupe Crédit Agricole

Ma vie. Ma ville. Ma banque.

GROUPE CRÉDIT AGRICOLE

38 marques

L'Humain

145 000 collaborateurs

Une présence mondiale dans **47 pays**

120 Métiers

1er employeur Privé en France

Résultats

1er gestionnaire d'actifs européen

1er investisseur dans l'immobilier en France

1er bancassureur en Europe

1er financeur de l'économie française

Présentation de LCL

Chiffres clés de LCL

Ma vie. Ma ville. Ma banque.

1863
Création
du Crédit Lyonnais

2003
Intégration du groupe
Crédit Agricole

2005
Naissance
de la marque LCL

17 300
collaborateurs

Plus de
1 700
implantations commerciales
(majoritairement en zone urbaine)

6 millions
de clients particuliers

354 000
clients professionnels

30 000
clients entreprises
et institutionnels

210 500
clients Banque Privée

1

Présentation de LCL

2

Présentation des activités de RCP\Modélisation

3

Présentation des grandes étapes de construction d'un score

4

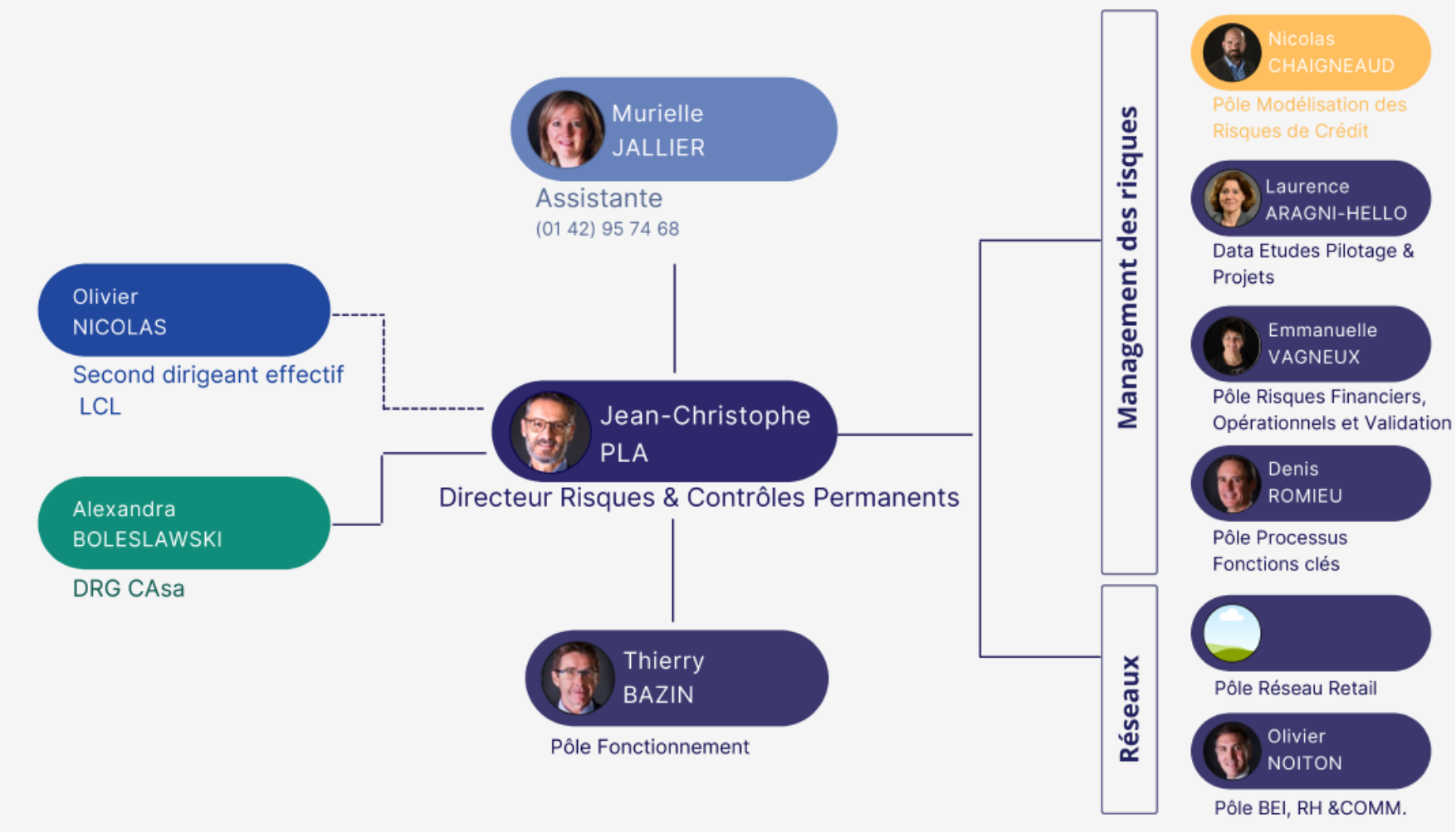
Présentation du projet

5

Ateliers

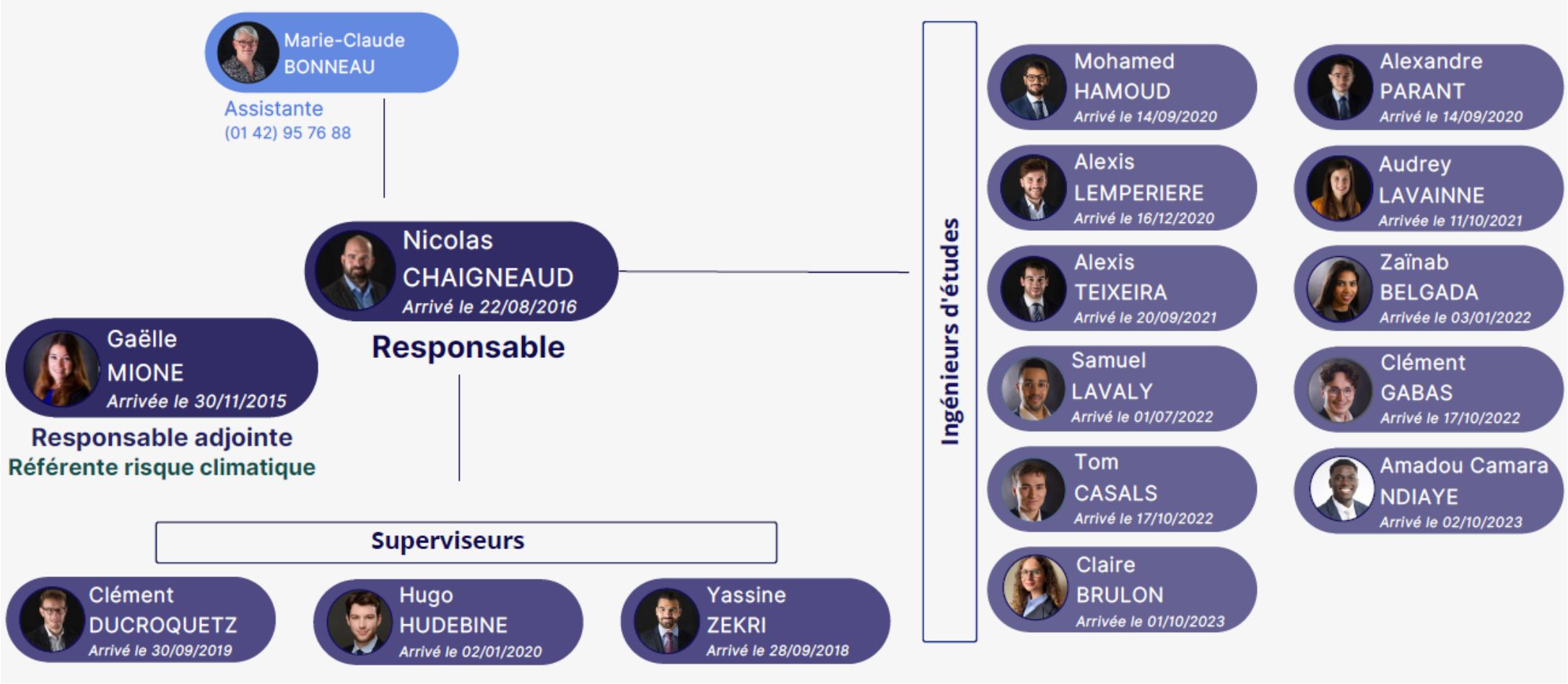
Présentation de RCP\Modélisation

Direction des Risques et Contrôles Permanents



Présentation de RCP\Modélisation

Organigramme



Présentation de RCP\Modélisation

Activités

Prudentiel - Bâle IV

Estimation des paramètres bâlois (e.g. probabilité de défaut, perte en cas de défaut) contribuant à la correcte maîtrise des fonds propres associés au risque de crédit

Financier/Comptable

Estimation des paramètres de provisionnement (IFRS 9, provisions individuelles statistiques) contribuant à la correcte maîtrise du risque de crédit

Scores

Construction, suivi et mise à jour des scores/outils d'aide à la décision utilisés par le Réseau (e.g. scores immobiliers, score d'octroi et pré-attribution aux professionnels, anticipation des risques, Aide à la Décision du Jour)

RCP

Modélisation

Reporting réglementaire

Mise en œuvre de nouveaux concepts réglementaires (e.g. restructuration pour risque, nouvelle définition du défaut, Bâle IV) et réalisation des exercices annuels de stress-tests

Risques climatiques

Intégration du risque climatique dans la gestion des risques : participation au guide BCE, mesure de l'empreinte carbone de nos financements (NZBA), mesure du risque physique, réalisation des stress test climatiques

1

Présentation de LCL

2

Présentation des activités de RCP\Modélisation

3

Présentation des grandes étapes de construction d'un score

4

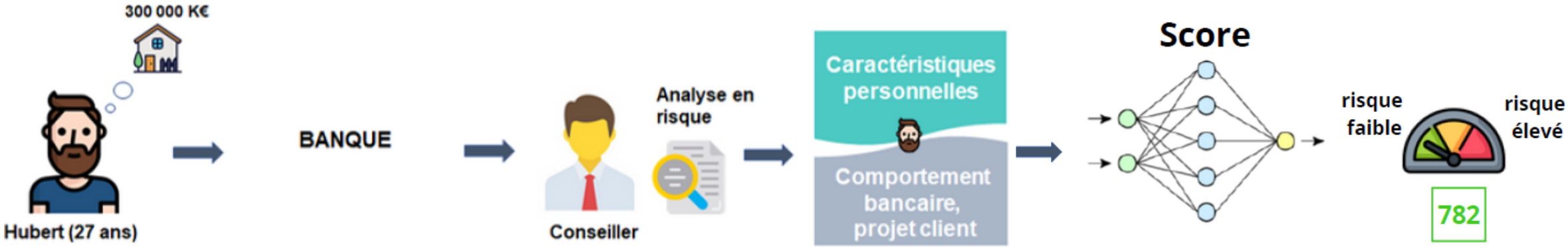
Présentation du projet

5

Ateliers

Les étapes de construction d'un score

Principe d'un score



Un score est un système de points, construit à partir des modèles statistiques, qui permet de noter les clients en fonction de leur risque présumé.

Sur la base d'une analyse des données signalétiques clients (âge, salaire, etc.) et de leur comportement bancaire (impayés, défauts, litiges, etc.), le score permet de regrouper les clients avec des caractéristiques similaires en termes de risque, en classes (appelées classes homogènes de risque).

Cela permet de prédire le niveau de risque d'une nouvelle demande de financement.

Les étapes de construction d'un score

Le scoring en banque

Que cherche t-on à prédire ?

Scoring

Appétence

Attrition

Marketing

Risques

Recouvrement

- notation ponctuelle
- prioriser les comptes en recouvrement
- client connus
- variables signalétiques/comportementales
- complétés par des règles experts

Réglementaires

Scores comportementaux

- notation récurrente
- clients connus
- variables comportementales
- calibrage de la PD
- utilisation transverse dans la banque

Octroi

Scores d'acceptation

- notation ponctuelle
- octroi de crédit et délégation
- clients connus ou prospects
- variables signalétiques et/ou financières
- peuvent être complétés par des règles experts
- score spécifique à un usage
- pas de données sur les refusés

Mais aussi : détection de la fraude, cas d'usage innovants, ...

Les étapes de construction d'un score

Exemples de score chez RCP\Modélisation

Score immobilier SE-BEST

Score immobilier Système Expert BEST (Banking Expert System Tool)

Aide à la décision à l'octroi d'un crédit immobilier selon le segment du client (Prospects ou Clients Connus) et selon le type d'investissement (locatif ou résidentiel).

NS PRO

Note de Signature Professionnel

Aide à la décision à l'octroi d'un crédit professionnel selon le segment du client : Prospects/Tiers Récents ou Clients Connus.

Scores CA-CF

Aide à la décision à l'octroi d'un crédit CA-CF pour les clients connus et prospects.

Early Warning Corporate

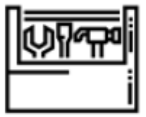
Système d'aide à la renotation potentielle des tiers Corporate en se basant sur les informations permettant d'anticiper le futur passage en sensibles de de ces derniers

ADJ PAR

Aide à la Décision du Jour Particuliers

Aide à décision quotidienne, paiement ou refus des mouvements effaçables (effets domiciliés, prélèvements LCL et hors LCL, chèques d'un montant supérieur à 15,24€) sur les comptes (Dépôts à Vue) en anomalie.

Pré-attribution des Professionnels



Permet aux conseillers d'améliorer l'équipement des clients professionnels en matière de crédits (Court Terme, Moyen Terme et Crédit Bail Mobilier) et de fidéliser les clients et prospects en étant pro actif.

Les étapes de construction d'un score

Quel est l'objectif ?

1. Définition des objectifs (8%)

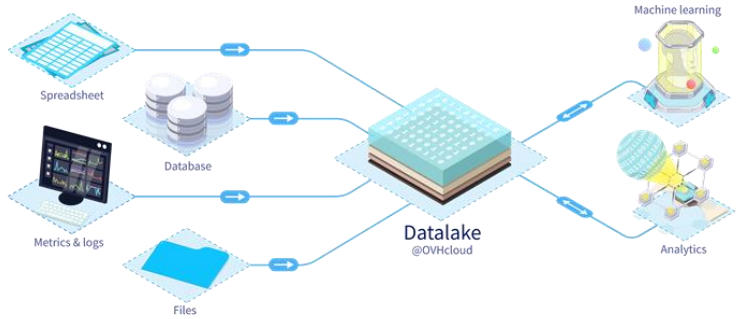
- **Intérêts du score, résultats attendus et population cible**
  - Nouveau score ou refonte ?
  - Problématique
  - Gains attendus
  - Utilisation opérationnelle

- **Données à modéliser**
  - Critère cible : défaut, nombre d'impayés, etc ?
  - Historique de données

- **Planification du projet**
  - Faisabilité du projet (volumétrie, historique, etc ?)
  - Deadlines
  - Réunions d'échanges

Les étapes de construction d'un score

Quelles sont les données disponibles ?



2. Inventaire et collecte des données (10%)

- **Cartographie des données**
  - Données internes
  - Données externes
  - Données légales, fiables, interprétables, implémentables dans le SI, disponibles dans le temps, etc.

- **Type de données**
  - Quelle granularité (mensuel, trimestriel, annuel ?)
  - Quelle agrégation (données de tous les emprunteurs, de l'emprunteur principale, du contrat, etc ?)

- **Collecte des données**
  - Import des données
  - Clé de jointure
  - Temps de traitement

Les étapes de construction d'un score

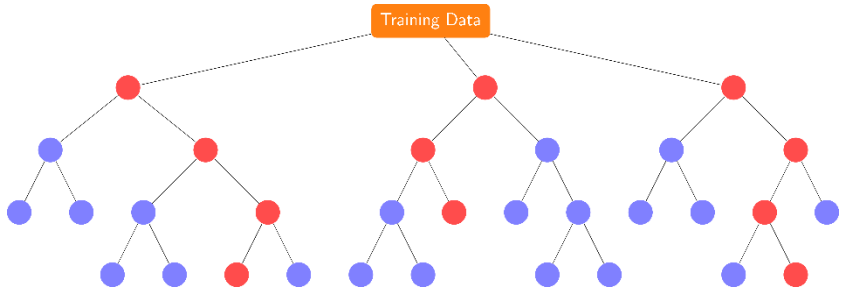
A quoi ressemblent mes données ?

3. Exploration et préparation des données (28%)

- **Analyses univariées**
  - Fréquences
  - Distributions
- **Statistiques descriptives**
  - Séries temporelles (évolution du taux de défaut)
  - % de valeurs manquants
  - Détection des valeurs extrêmes et aberrantes
- **Analyses bivariées**
  - Variables explicatives entre elles (corrélation, création de nouvelles variables)
  - Avec la variable cible
- **Travail sur les données**
  - Création de nouvelles variables (ratios, évolution temporelle, etc.)
  - Réduction du nombre de variables (selon la qualité, la corrélation avec la variable cible, la pertinence, etc.)
  - Fiabilisation des données (gestion valeurs manquantes, valeurs extrêmes, etc.)
  - Discrétisation/regroupement de modalité.

Les étapes de construction d'un score

Quelles techniques utiliser ?



4. Segmentation et élaboration du modèle (25%)

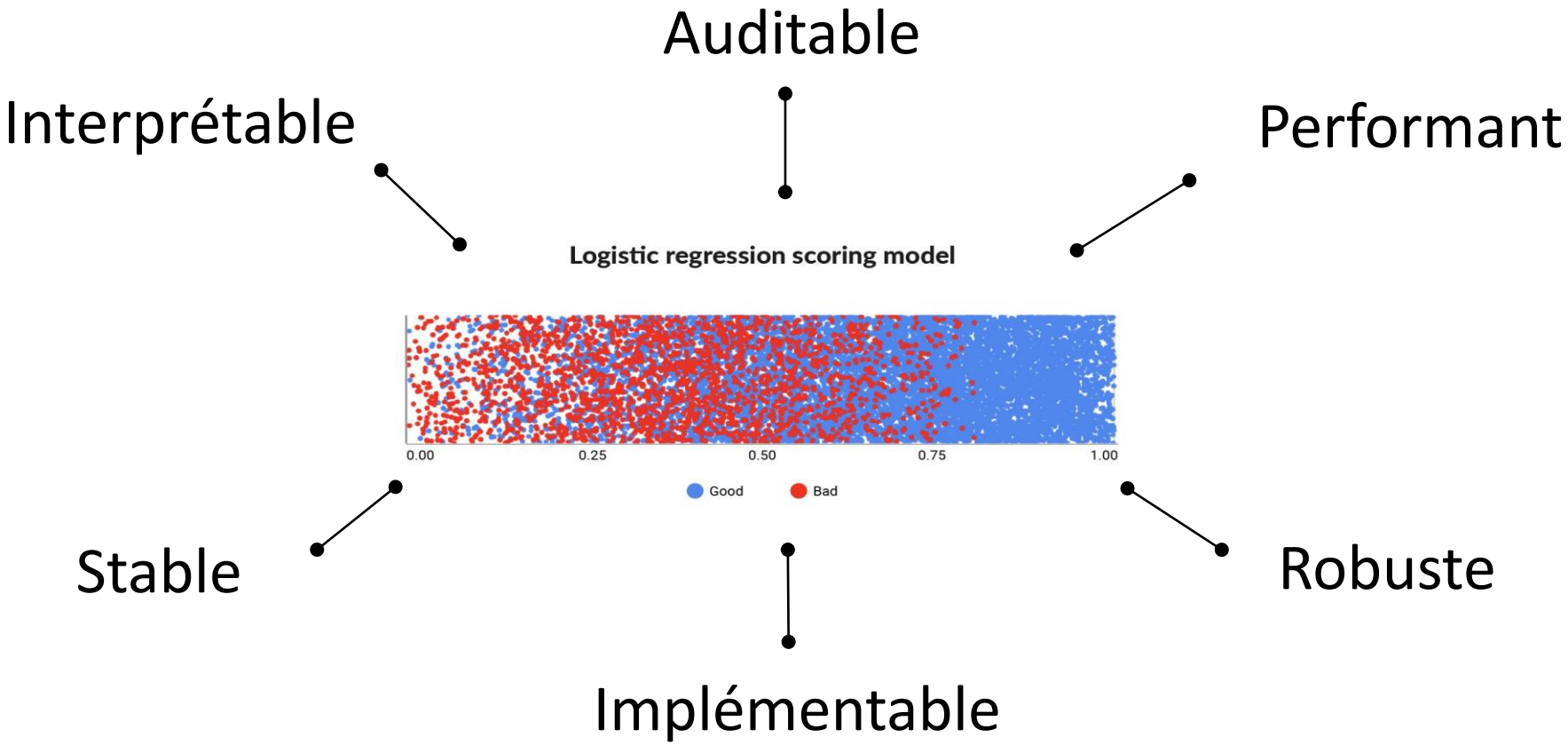
- **Classification (si population suffisante)**
  - Supervisée, non supervisée, règles expertes
  - Nombre de segments

- **Modèles prédictifs**
  - Choix du modèle (régression logistique, arbres, deep learning, etc.)
  - Elaboration de plusieurs modèles (utilisation de benchmark)
  - Elaboration du modèle pas à pas pour garantir la performance

- **Performance**
  - Choix entre plusieurs modèles
  - Indicateurs statistiques (R^2, indice de Gini, courbe ROC)

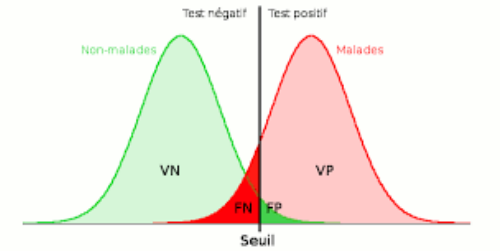
Les étapes de construction d'un score

Pourquoi la régression reste dominante ?



Les étapes de construction d'un score

Mon modèle est-il performant ?



5. Validation du modèle (12%)

- **Validation du modèle**
  - Echantillon test, out-of-time, bootstrap, etc.

- **Validation métier**
  - Pertinence métier des variables explicatives, des découpages
  - Soumission de cas réels
  - Ajout de forçages, règles expertes en aval ?

- **Validation méthodologie**
  - Validation interne
  - Inspection générale
  - Homologation BCE

- **Impacts opérationnels et financiers**

Les étapes de construction d'un score

Comment utiliser mon score opérationnellement ?

6. Déploiement du modèle (7%)

- **Implémentation dans le SI**
  - Accompagnement de l'IT
  - Recette métier
  - Comitologie

- **Formation des utilisateurs**

7. Suivi des modèles et améliorations (10%)

- **Backtesting réguliers**
  - Analyse statistique
  - Communication des résultats
  - Plans d'actions (ajout de règles, remplacement de variables, recalibrage, etc.)

1

Présentation de LCL

2

Présentation des activités de RCP\Modélisation

3

Présentation des grandes étapes de construction d’un score

4

Présentation du projet

5

Ateliers

Présentation du projet

Une augmentation du risque sur les Professionnels

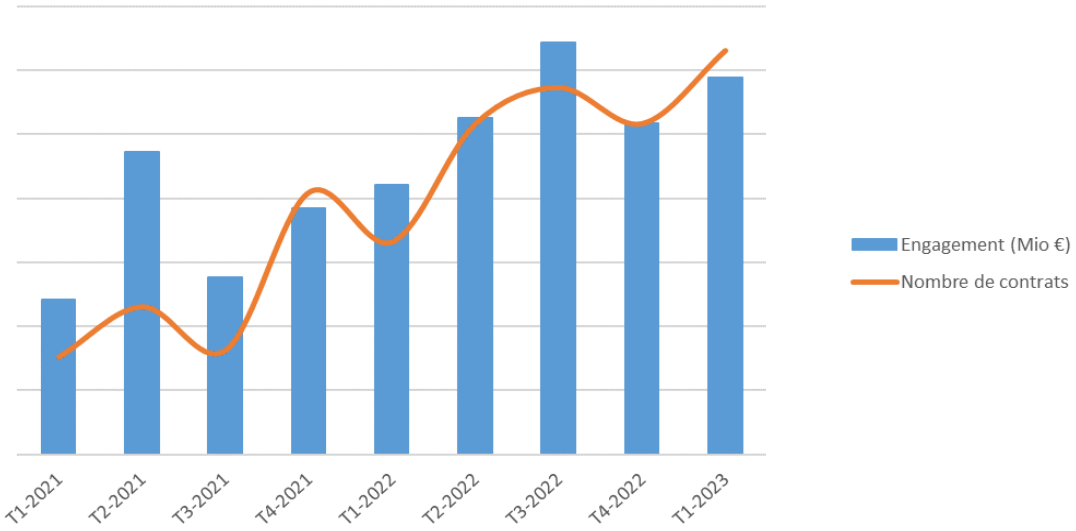
Contexte

Le niveau des défaillances d'entreprises en France progresse (au T1-2023 : 46.700, cumul 12 mois) sans atteindre toutefois les niveaux pré-Covid (51.000) mais avec des situations contrastées selon les segments et des prévisions dépassant le seuil de 55.000 à fin de 2023.

Sur le début 2023 on relève une accélération, avec 14 317 défaillances sur le premier trimestre, en hausse de 43,6% vs le 1er trimestre de l'année 2022 et dépassant le niveau d'avant crise.

Chez LCL - quelques chiffres

Evolution des entrées en défaut trimestrielles depuis 2021
Périmètre : hors Interfimo et SCI



Lancement des travaux

Pourquoi ?

Dans un contexte d'augmentation du risque sur la clientèle professionnelle, l'objectif est de détecter en avance de phase des poches de clients susceptibles de faire l'objet d'impayés / de faillite.

Pour quoi ?

Les intérêts sont multiples :

1. Identifier les symptômes entraînant l'apparition de difficultés financières dans un futur proche ;
2. Cibler et accompagner les clients en avance de phase ;
3. D'un point de vue risque, suivre ces poches de clients dans le temps et optimiser le provisionnement des encours.

Comment ?

La construction d'un score d'anticipation du risque est en cours côté RCP\Modélisation permettant de discriminer au mieux les clients selon le critère cible.

Présentation du projet

Défaut bâlois

Arriérés de paiement

- Le défaut est déclenché en cas de présence d'arriérés de paiement (dépassement et/ou impayés) significatifs (dépassement du seuil de signification) pendant plus de 90 jours consécutifs.

Composantes du seuil de signification :

Seuil Absolu
Arriérés du débiteur
> 100 € Retail
> 500 € Hors-Retail

+

Seuil Relatif
Arriérés / Engagements de bilan
> 1%

Seuils franchis
> 90 jours consécutifs

- Le défaut a lieu si le seuil absolu **et** le seuil relatif sont dépassés simultanément pendant plus de 90 jours consécutifs.

Unlikeliness to pay (UTP)

- La présence d'un signe de probable absence de paiement (*i.e.* Unlikeliness To Pay – UTP) déclenche également le passage en défaut.
- Chez LCL, 8 UTP ont été définies (fraude, surendetemment, procédures collectives, *etc.*)

Sortie du défaut

3 mois
dès qu'il n'y a plus d'élément déclencheur du défaut

- Si pas incident :

1 → 2 → 3 ✓ **Retour en sain**

- Si incident en fin de période (seuils franchis) :

1 → 2 → 3 ✗ **Prolongement au jour le jour** jusqu'à régularisation

- Si incident au cours de la période (seuils franchis pendant plus de 30 jours) :

1 ✗ 1 → 2 → 3 **Réinitialisation** de la période de surveillance

Présentation du projet

Construction du score testé par LCL

Périmètre de scoring

Périmètre initial

365 470 clients professionnels non en défaut à date et notés par le modèle IRPRO

Filtres d'exclusion

1. **Clients sans engagement à date**
2. **Périmètres spécifiques** (Interfimo, SCI patrimoniales, clientèles BP et GF)
3. **Clients avec arriérés significatifs à date**

Périmètre final

- **109 926 clients** (soit 30% du périmètre initial)

Critère cible testé par LCL

Définition d'une **variable cible hybride à horizon 3 mois** (entre m+1 et m+3), comprenant au moins l'un des critères suivants :

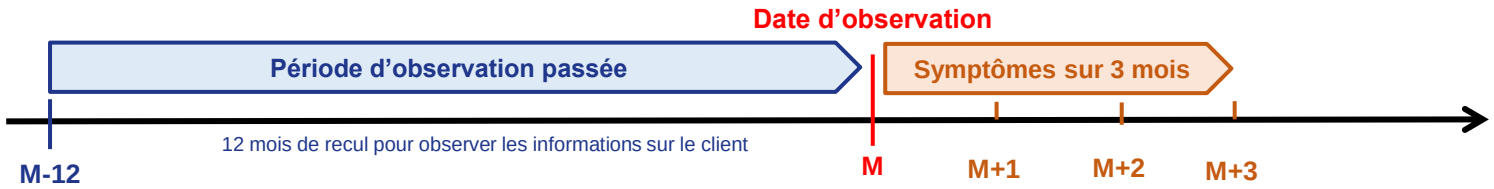
- 1 impayé significatif ;
- Dépassements significatifs > 20 jours ;
- Passage en procédure collective (liquidation judiciaire, redressement judiciaire et procédure de sauvegarde).

Soit en moyenne (sur la période 202202-202206), **4007 clients** concernés mensuellement, pour un taux de variable cible de l'ordre de **3,92%**.

Une étude quantitative sur la période en question a permis de mettre en évidence, pour le périmètre de clients identifiés dans notre variable cible avec impayé et/ou dépassements, que 44% d'entre eux atteignent les 90 jours d'arriérés significatifs dans les mois qui suivent.

Méthodologie

Observation des données sur une fenêtre glissante de 13 mois (à date et sur les 12 derniers mois), afin d'anticiper l'apparition de symptômes dans les trois mois qui suivent :



L'historique final retenu pour le calibrage du score est l'année 2022, avec donc une fenêtre d'observation totale de 27 mois.

Présentation du projet

Définitions challengers

Définition 1

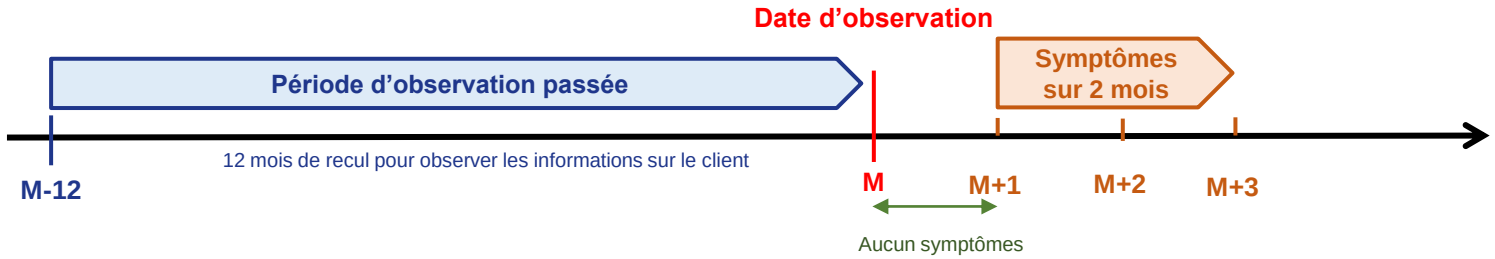
Semblable à la définition du score LCL, mais un nombre de jours de dépassement supérieur à 10 jours ;

- 1 impayé significatif ;
- Dépassements significatifs > 10 jours ;
- Passage en procédure collective (liquidation judiciaire, redressement judiciaire et procédure de sauvegarde).

Nom de la variable dans la base : **Cible_1**

Définition 2

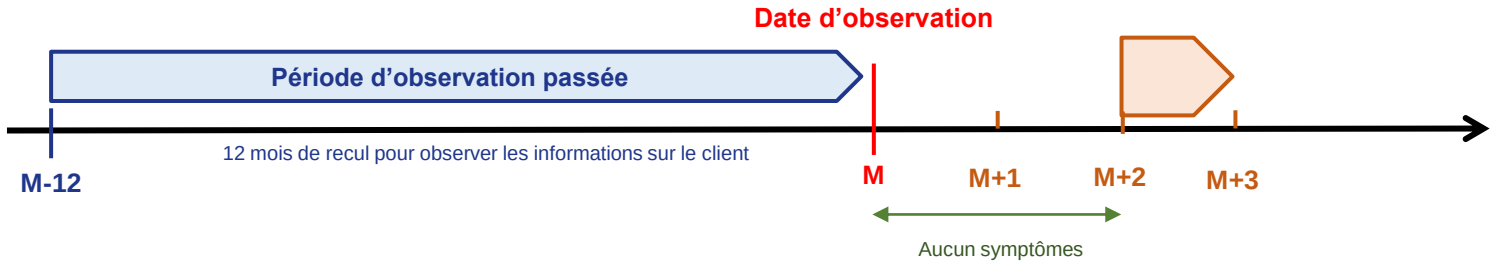
Définition d'une **variable cible hybride**, sans incident à m+1, comprenant au moins l'un des 3 critères sur m+2 et m+3.



Nom de la variable dans la base : **Cible_2**

Définition 3

Définition d'une **variable cible hybride**, sans incident à m+1 et M+2, comprenant au moins l'un des 3 critères sur m+3 :



Nom de la variable dans la base : **Cible_ 3**

Aucun symptômes

1

Présentation de LCL

2

Présentation des activités de RCP\Modélisation

3

Présentation des grandes étapes de construction d’un score

4

Présentation du projet

5

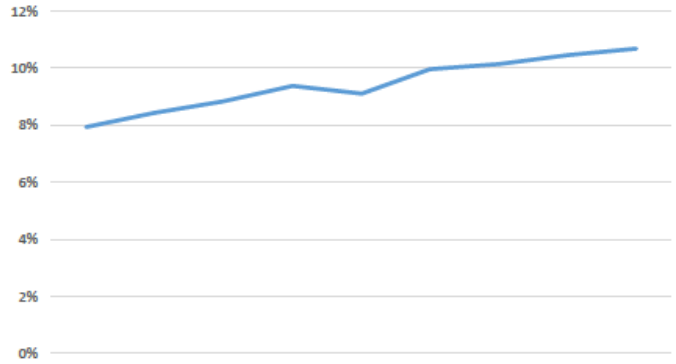
Ateliers

A quoi ressemblent mes données ?

Que contient ma base de données ?

Travail :

- Prise en main du dictionnaire de données ;
- Identification de la variable à expliquer et analyses ;
- Identification des variables quantitatives et qualitatives ;
- Tris à plat, analyse univariée/bivariée, graphique temporel (dont critère cible).



Quelle est la qualité de mes données?

Travail : détection

- des valeurs manquantes,
- des valeurs aberrantes,
- des valeurs extrêmes (« outliers »),
- des doublons

Variable explicative	Variable cible	Q1	Médiane	Moyenne	Q3	Valeurs man-quantes
Encours du compte de dépôt	0 1	65.45 -2945.57	1844.31 186.46	24689.30 -287.33	10379.95 2113.00	0
Nombre de jours débiteurs consécutifs en dépassement de l'autorisation	0 1	0 0	0 0	2.63 10.83	0 10.00	4
Nombre de jours d'arriérés	0 1	0 0	0 0	1.28 8.20	0 5.00	0

Puis-je créer de nouveaux indicateurs ?

Travail :

- Calcul d'évolution temporelles ;
- Combinaison linéaires ;
- Recodage de variables ;
- Remplacement de date par des durées.

Comment sélectionner mes variables ?

Comment échantilloner ?

Travail : échantillonage stratifié (70%-30%) – définir les règles de stratification

- ✓ vérifier la représentativité des échantillons d'apprentissage et de test vis-à-vis de la base d'analyse.

➡ La base d'apprentissage (70%) est utilisée par la suite pour la sélection des variables et l'estimation du modèle.

Comment étudier les liaisons entre les variables ?

Travail : détecter les variables candidates potentiellement discriminantes et identifier les variables explicatives trop liées entre elles (des croisements de variables peuvent potentiellement être déduis des variables explicatives trop liées).

		Variables	
		Quantitatives	**Qualitatives**
Variables	**Quantitatives**	Coefficients de corrélation de Spearman et de Pearson	Kruskal-Wallis
	Qualitatives	Kruskal-Wallis	T de Tshuprow V de Cramer

Des méthodes de type lasso, arbres de décisions, …., sont également possibles.

Comment sélectionner mes variables ?

Quelles sont les variables qualitatives les plus discriminantes?

Travail : Etudier les liaisons des variables explicatives qualitatives avec la variable cible

- Utilisation du Chi2 ou du V de CRAMER (intègre l'effectif et le nombre de degré de liberté)
- Utilisation du T de TSHUPROW (permet de comparer les différents découpages entre eux)

➡ Classement des variables selon leur lien avec le critère cible

Exemple de sorties :

Variable	Stat. du χ^2	p-valeur	V de Cramer	T de Tschuprow
Code sous-produit	Effectif <5	-	0.012782	0.012782
Situation familiale	61.2685	<0,0001	0.0451	0.030187
Nature du bien	Effectif <5	-	0.010994	0.010994
Type de contrat client	15.8776	0.0032	0.0230	0.016249
Nombre d'enfants	Effectif <5	-	0.028014	0.028014
Présence co-emprunteur	101.7979	<.0001	0.0582	0.058186
Statut résidentiel	282.3999	<.0001	0.0969	0.064809
CSP client	84.5546	<.0001	0.0530	0.023053

Quelles sont les variables quantitatives les plus discriminantes?

Travail : Etudier les liaisons et détecter les variables discriminantes continues les plus corrélées avec la variable cible

- Kruskall-Wallis (statistique de rang, sensible au degré de liberté)

➡ Classement des variables selon leur lien avec le critère cible

Exemple de sorties :

Variable	Stat. test KW	p-valeur
Revenu du ménage	125,7829	<,0001
Ancienneté embauche client	90,9019	<,0001
Durée dans le logement actuel	37,8669	<,0001
Montant du loyer	35,3433	<,0001
Revenu principal client	22,3162	<,0001
Age client	20,6314	<,0001
Ratio Montant du crédit accordé/Revenu ménage	16,9985	<,0001
Montant du crédit accordé	7,1315	0,0076
Nombre d'échéances du crédit	6,625	0,0101
Ratio Loyer/Revenu ménage	0,0526	0,8185

Comment sélectionner/transformer mes variables ?

Les variables explicatives sont-elles corrélées ?

Travail : Détecter les liaisons linéaires entre les variables explicatives

- PEARSON / SPEARMAN pour les variables quantitatives (à surveiller > 0.7, dangereuse > 0.8, inacceptable > 0.9)
- V de CRAMER pour les variables qualitatives (inacceptable > 0,7)

➡ Réflexion sur les croisements possibles à réaliser entre les variables très corrélées

Exemple de sorties :

	Situation familiale	Présence coemprunteur	Statut résidentiel	CSP	Nombre d'enfants
Situation familiale					
Présence co-emprunteur	0,5511				
Statut résidentiel	0,1217	0,2226			
CSP	0,156	0,1604	0,1294		
Nombre d'enfants	0.1645	0.2311	0.0754	0.1347	

	Age client	Revenu ménage	Revenu client	Crédit acc./ rev. ménage	Montant loyer	Anc. embauche client	Anc. emménagement client	Crédit accordé	Nombre échéances
Age client									
Revenu ménage	0.01770								
Revenu principal client	0.01369	0.52470							
Crédit accordé/revenu ménage	0.00272	-0.14360	-0.04900						
Montant loyer	-0.16347	0.36611	0.30024	-0.13111					
Ancienneté embauche client	0.25375	0.10942	0.13631	-0.02842	0.01555				
Ancienneté emménagement client	0.41860	-0.00196	-0.04211	0.02235	-0.21435	0.21682			
Montant crédit accordé	0.01660	0.27289	0.24271	0.60319	0.07735	0.04518	0.02363		
Nombre d'échéances	-0.04354	0.07210	0.07302	0.39183	0.06459	-0.00314	-0.05110	0.42529	

Comment transformer mes variables ?

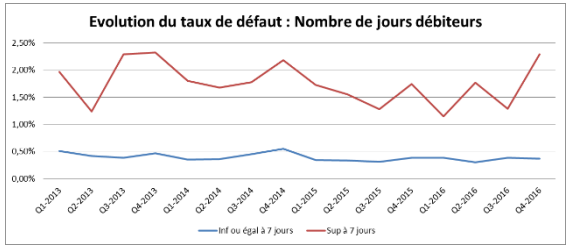
Comment discrétiser les variables continues ?

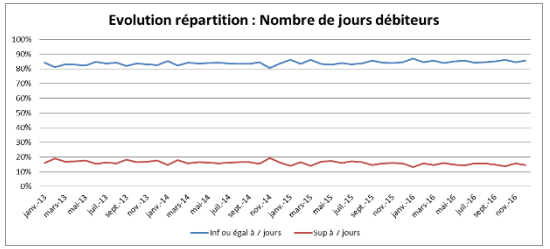
- Limiter l'impact des données aberrantes, prendre en compte des effets non linéaires ou des interactions (pour améliorer la lisibilité de la grille de score finale)
- Gérer les valeurs manquantes et les valeurs extrêmes

- **APPROCHE GRAPHIQUE** : réaliser une courbe de densité conditionnelles ou histogramme de la distribution de la variable quantitative en fonction du phénomène à prédire : l'analyste décide des bornes du découpage en fonction du graphique

- **APPROCHE STATISTIQUE** : découpage pas à pas
  1. Découper en N classes ;
  2. Regrouper les modalités (selon le taux de risques, stabilités, cohérence métier)
  3. Réaliser des **transformations de 2 à 5 modalités** par variable en réunissant **au moins environ 5% de la population**
  4. Retenir uniquement les transformations dont les modalités sont **séparées d'au moins 30% en termes de risque** (écarts relatifs entre les taux de défaut)
  5. Vérifier les **stabilités en volume et en risque** des transformations ainsi retenues (choix d'un pas de temps cohérent)
  6. Calculer les indicateurs statistiques T de Tschuprow et V de Cramer pour chaque transformation.

Exemple de sorties :





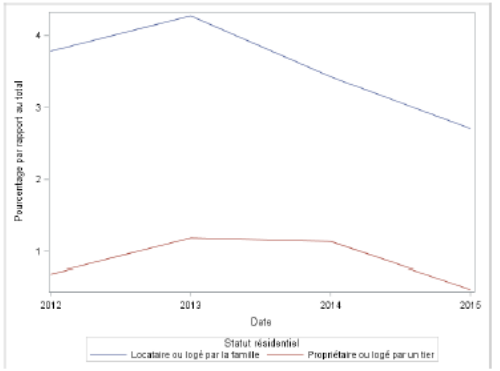
Comment transformer mes variables ?

Comment regrouper les variables qualitatives ?

- Limiter le nombre de modalités, recoder et regrouper les modalités (notamment si faible effectif par modalité)

- **APPROCHE STATISTIQUE** : découpage pas à pas
  1. Recoder les modalités avec des noms trop long ou imprécis
  2. Regrouper les modalités (selon le taux de risques, stabilités, cohérence métier)
  3. Réaliser des **transformations de 2 à 5 modalités** par variable en réunissant **au moins environ 5% de la population**
  4. Retenir uniquement les transformations dont les modalités sont **séparées d'au moins 30% en termes de risque** (écarts relatifs entre les taux de défaut)
  5. Vérifier les stabilités en volume et en risque des transformations ainsi retenues (choix d'un pas de temps cohérent)
  6. Calculer les indicateurs statistiques T de Tschuprow et V de Cramer pour chaque transformation.

Exemple de sorties :



Statut résidentiel	Logé par l'administration	Logé par l'employeur	Logé par la famille	Logé par le concubin	Locataire	Propriétaire
Pourcentage de la base (en %)	0,41	1,44	3,05	2,44	32,32	60,33
Proportion de défaut dans la classe (en %)	0,00	2,07	3,70	3,81	3,57	0,86

Comment définir mes variables finales à tester ?

Quelles sont les variables à retenir pour la régression ?

Travail : Etudier les liens entre les variables explicatives et la variable cible ainsi qu'entre les variables explicatives entre elles *post* transformation

- V de CRAMER pour les variables qualitatives (inacceptable > 0,7)

➡ Limiter le nombres de variables à tester dans le modèle

Exemple de sorties :

	Statut résidentiel	Revenu ménage	Présence co-emprunteur	Ancienneté embauche	Montant loyer	Situation familiale	CSP client	Ancienneté emménagement	Age client	Nombre d'enfants	Montant crédit/ Revenu ménage
Statut résidentiel											
Revenu ménage	0,319										
Présence co-emprunteur	0,2049	0,3836									
Ancienneté embauche	0,1562	0,0976	0,0385								
Montant loyer	0,3659	0,3618	0,2692	0,0487							
Situation familiale	0,2073	0,2879	0,5268	0,0934	0,1468						
CSP client	0,1617	0,2428	0,0543	0,0874	0,2447	0,1303					
Ancienneté emménagement	0,1586	0,0635	0,0588	0,2135	0,258	0,1761	0,2361				
Age client	0,1503	0,1069	0,076	0,2622	0,2407	0,2114	0,5031	0,3654			
Nombre d'enfants	0,0181	0,1329	0,0809	0,0511	0,1069	0,1811	0,1202	0,0842	0,2022		
Montant crédit/ revenu ménage	0,0001	0,1618	0,002	0,0335	0,1174	0,1099	0,0681	0,0276	0,0428	-0,0345	

Comment estimer mon modèle ?

1. Démarche préconisée : processus itératif en partant d'une liste vide et on intègre une à une les variables (possibilité de comparer avec ce que donne une stepwise, backward, forward) avec vérification de plusieurs conditions.

2. Comparaison des performances et de la qualité si plusieurs modèles (AIC, BIC, R^2, etc.);

3. Analyse des performances du modèles sélectionné (Gini, courbes de densité conditionnelles, courbe ROC, etc.) ;

4. Construction de la grille de score (calcul des pondérations, calcul des contributions de chacune des variables) ;

5. Construction de l'échelle de risque (segmentation en CHR).

Comment construire mon modèle ?

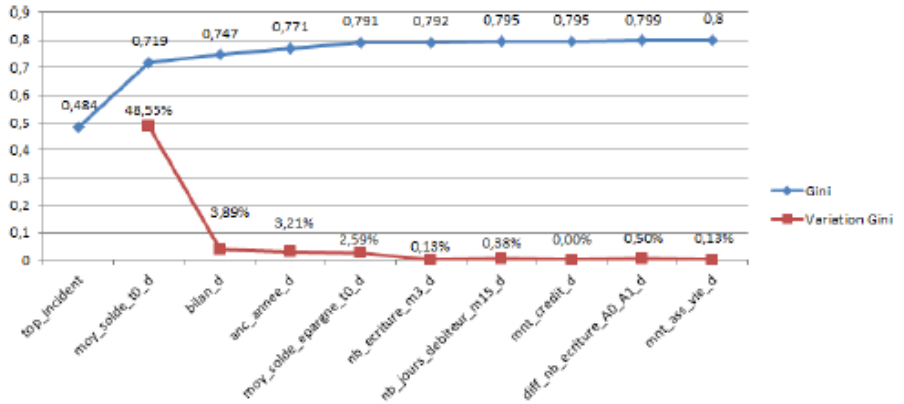
1. Intégration des variables au modèle de régression logistique

Travail :

- Intégrer de manière itérative dans le modèle la variable explicative candidate à laquelle est associée la valeur de la statistique du V de Cramer la plus élevée ;
- Observer la significativité (p-values) de la variable et des modalités de la variable ;
- Vérifier la cohérence du (des) coefficient(s) de la variable vis-à-vis du niveau de risque et de l'expertise métier (en particulier la cohérence des coefficients vis-à-vis du critère à modéliser, le signe du coefficient doit être en cohérence avec l'effet attendu positif ou négatif) ;
- Mesurer l'apport d'information portée par cette variable. Effectuer notamment un suivi de l'apport marginal de chaque variable en fonction de l'indice de Gini ;
- Les odds ratio (OR) estimés par le modèle sont tous bien positifs et leurs intervalles de confiance ne contiennent pas 1, nous pouvons donc les interpréter : ils sont significatifs.

Règles de décision :

- Conserver la variable au sein des variables du score si, d'une part, la significativité des coefficients de ses modalités est supérieure au seuil de 5 % (l'apport d'information portée par cette variable est non nul), et si, d'autre part, l'information qu'elle porte ne peut pas être reliée de manière excessive à celle portée par l'une des variables déjà retenues (contrôle de l'intensité des liaisons) ;
- Reporter l'intégration de la variable si elle n'apporte que peu d'information (significativité des coefficients de ses modalités compris entre 5 et 10 %, apport faible d'information). Cette variable sera à "re-tester" plus tard afin de « peaufiner » éventuellement le modèle de score ;
- Exclure la variable du score si tous les coefficients de ses modalités sont non significatifs (au seuil de 10 %) (l'apport d'information est nul) ;
- Si la variable est significative au seuil de 5 % mais que l'une de ses modalités ne l'est pas, il convient de tester un découpage différent ou de regrouper cette modalité avec une autre modalité ayant un coefficient proche ou un taux de défaut proche,
  tout en veillant à conserver la cohérence métier.



Comment sélectionner mon modèle ?

2. Comparaison des performances

Travail :

- Dans le cas ou plusieurs modèles sont testés, il n'est pas nécessaire de procéder à l'étude des capacités prédictives des scores estimés pour pouvoir comparer plusieurs modèles concurrents. Le calcul des critères d'information permet en effet de classer les modèles en rapportant leur qualité d'ajustement à leur degré de parcimonie.

- Les critères d'information sont ainsi particulièrement utiles dans une optique de mesure d'apport d'information portée par une variable. Les plus utilisés sont **l'AIC** (Akaike Information Criterion) et le **BIC** (Bayesian Information Criterion) / Schwartz qui diffère de l'AIC de par la pénalité associée au nombre de paramètres estimés.

- L'utilisation conjointe de ces deux critères permet in fine de disposer d'un premier classement entre plusieurs modèles cohérents et valides. Au final, on préférera les modèles possédant un critère AIC (resp. BIC) faible.

3. Qualité du modèle

Travail :

- Les qualités prédictives d'un modèle peuvent aussi s'appréhender via l'utilisation du R^2.

- L'idée sous-jacente à cet indicateur est de considérer le pouvoir explicatif du modèle testé au regard d'un modèle théorique parfait, ce qui permet de disposer in fine d'un indicateur compris entre 0 et 1 : plus le R^2 est élevé, meilleur est le modèle.

Comment sélectionner mon modèle ?

Test de significativité globale + significativité des coefficients

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22282.8746	9	<.0001
Score	23193.7846	9	<.0001
Wald	18068.9501	9	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
IRISQUE_PRO_fin	2	3462.5242	<.0001
NBJDEPDP_fin	1	1071.4142	<.0001
MNTECSCPTDPT_fin	2	638.5107	<.0001
MNT_DEP_TIERS	1	888.5399	<.0001
NBRETCAR_M12_fin	2	1648.2664	<.0001
ANCIENNETE_LCL_fin	1	939.0396	<.0001

Comment sélectionner mon modèle ?

Test de significativité des modalité des variables

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.2251	0.0183	151.5323	<.0001
IRISQUE_PRO_fin	1<=IRPRO<=5	1	1.3932	0.0238	3414.3860	<.0001
IRISQUE_PRO_fin	6<=IRPRO<=7	1	0.4905	0.0178	758.6008	<.0001
IRISQUE_PRO_fin	>7 ou Non renseigné	0	0	.	.	.
NBJDEPDP_fin	<=7j	1	0.6088	0.0186	1071.4142	<.0001
NBJDEPDP_fin	>7j	0	0	.	.	.
MNTECSCPTDPT_fin	<=5000€	1	0.2188	0.0163	180.4353	<.0001
MNTECSCPTDPT_fin	>5000€	1	0.5771	0.0229	634.0139	<.0001
MNTECSCPTDPT_fin	<=0€	0	0	.	.	.
MNT_DEP_TIERS	Pas de dépassement	1	0.4557	0.0153	888.5399	<.0001
MNT_DEP_TIERS	Depassement > 0€	0	0	.	.	.
NBRETCAR_M12_fin	<= 1 retrait	1	0.6013	0.0149	1637.3497	<.0001
NBRETCAR_M12_fin	entre 2 et 4 retaits	1	0.2636	0.0227	134.9024	<.0001
NBRETCAR_M12_fin	>4 retraits	0	0	.	.	.
ANCIENNETE_LCL_fin	+ de 10 ans	1	0.5173	0.0169	939.0396	<.0001
ANCIENNETE_LCL_fin	- de 10 ans	0	0	.	.	.

Comment sélectionner mon modèle ?

Performances

Association of Predicted Probabilities and Observed Responses

Percent Concordant	75.5	Somers' D	0.527
Percent Discordant	22.9	Gamma	0.535
Percent Tied	1.6	Tau-a	0.090
Pairs	6460107888	c	0.763

Gini

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	171849.66	149584.78
SC	171860.18	149690.03
-2 Log L	171847.66	149564.78

R-Square	0.0778	Max-rescaled R-Square	0.1675

R^2

Comment sélectionner mon modèle ?

Odds ratios

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
IRISQUE_PRO_fin	1<=IRPRO<=5 vs >7 ou Non renseigné	4.028	3.844	4.221
IRISQUE_PRO_fin	6<=IRPRO<=7 vs >7 ou Non renseigné	1.633	1.577	1.691
NBJDEPDP_fin	<=7j vs >7j	1.838	1.772	1.906
MNTECSCPTDPT_fin	<=5000€ vs <=0€	1.245	1.206	1.285
MNTECSCPTDPT_fin	>5000€ vs <=0€	1.781	1.703	1.863
MNT_DEP_TIERS	Pas de dépassement vs Depassement > 0€	1.577	1.531	1.625
NBRETCAR_M12_fin	<= 1 retrait vs >4 retraits	1.824	1.772	1.878
NBRETCAR_M12_fin	entre 2 et 4 retaits vs >4 retraits	1.302	1.245	1.361
ANCIENNETE_LCL_fin	+ de 10 ans vs - de 10 ans	1.678	1.623	1.734

Contrôler que 1 n'appartient pas aux intervalles

Ma vie. Ma ville. Ma banque.

Comment sélectionner mon modèle ?

Table de classification

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensitivity	Specificity	False POS	False NEG
0.540	249E3	0	25944	0	90.6	100.0	0.0	9.4	.
0.560	248E3	1617	24327	1465	90.6	99.4	6.2	8.9	47.5
0.580	248E3	1617	24327	1465	90.6	99.4	6.2	8.9	47.5
0.600	248E3	1617	24327	1465	90.6	99.4	6.2	8.9	47.5
0.620	246E3	2255	23689	2657	90.4	98.9	8.7	8.8	54.1
0.640	246E3	2255	23689	2657	90.4	98.9	8.7	8.8	54.1
0.660	246E3	2255	23689	2657	90.4	98.9	8.7	8.8	54.1
0.680	245E3	3012	22932	4279	90.1	98.3	11.6	8.6	58.7
0.700	24E4	5317	20627	9187	89.2	96.3	20.5	7.9	63.3
0.720	239E3	5433	20511	9851	89.0	96.0	20.9	7.9	64.5
0.740	238E3	5949	19995	11121	88.7	95.5	22.9	7.8	65.1
0.760	235E3	6986	18958	14300	87.9	94.3	26.9	7.5	67.2
0.780	234E3	7093	18851	14948	87.7	94.0	27.3	7.5	67.8
0.800	225E3	9481	16463	23718	85.4	90.5	36.5	6.8	71.4
0.820	213E3	11897	14047	36246	81.7	85.4	45.9	6.2	75.3
0.840	207E3	13243	12701	42372	80.0	83.0	51.0	5.8	76.2
0.860	201E3	14082	11862	48021	78.2	80.7	54.3	5.6	77.3
0.880	188E3	15920	10024	60547	74.3	75.7	61.4	5.1	79.2
0.900	166E3	18755	7189	83323	67.1	66.5	72.3	4.2	81.6
0.920	153E3	20110	5834	96155	62.9	61.4	77.5	3.7	82.7
0.940	127E3	22193	3751	122E3	54.2	50.9	85.5	2.9	84.6
0.960	101E3	23719	2225	148E3	45.5	40.7	91.4	2.1	86.2
0.980	36478	25364	580	213E3	22.5	14.6	97.8	1.6	89.3
1.000	0	25944	0	249E3	9.4	0.0	100.0	.	90.6

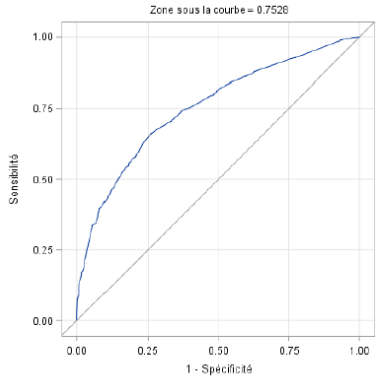
Comment vérifier que mon modèle est performant ?

4. Analyse des performances

Travail : Analyser les performances prédictives du scores sur l'échantillon de test

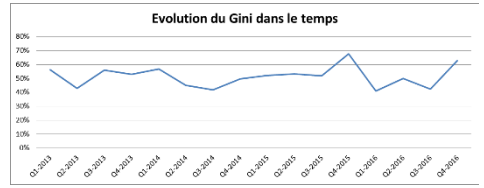
- L'analyse des performances prédictives du score obtenu via l'estimation d'un modèle est essentielle. De plus, la comparaison du pouvoir prédictif associé à plusieurs modèles concurrents permet d'effectuer un ranking de ceux-ci, facilitant le choix du score qui sera in fine retenu.

- Divers indicateurs standard permettent de quantifier les performances associées à un score :
  - **Courbes ROC** : courbe représentant l'arbitrage entre vrai et faux positifs
    - ✓ Comparer courbe sur échantillon test et apprentissage



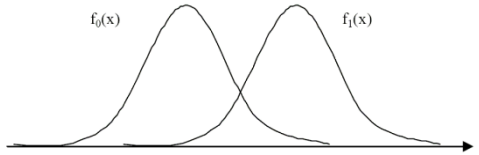
  - **Taux d'erreur de classement**

	Contreparties constatées saines	Contreparties constatées défaillantes
Contreparties prévues saines par le score au seuil s	A	B
Contreparties prévues défaillantes par le score au seuil s	C	D



  - **Indice de GINI** : indicateur compris entre 0 et 1 qui rend compte de la répartition d'une variable dans une population. Dans un score, on s'attend qu'il soit proche de 1, signe que les défauts sont concentrés dans les classes les plus risquées. Il est égal à 2 * AUC – 1.
    - ✓ Comparer courbe sur échantillon test et apprentissage.

  - **Courbes de densité conditionnelles** : Distribution des scores des individus conditionnellement au défaut. Plus les distributions sont éloignées, plus le score est discriminant.
    - ✓ Regarder sur échantillon test.



Comment rendre lisible et interprétable mon modèle ?

5. Comment construire la grille de score ?

Travail : Une fois la régression logistique effectuée, calculer les pondérations associées aux modalités des variables dans le but de créer une échelle de score allant de 0 à 1000. Un client avec un score élevé sera alors estimé comme peu risqué et, a contrario, un client avec un score faible sera considéré comme ayant une forte probabilité de tomber en défaut.

- **Calcul des pondération** : notons $c(j,i)$ et $SC(j,i)$ respectivement le coefficient du modèle et la pondération associés à la modalité i de la variable j, et αj=max [$c(j,i)$] le coefficient maximum pour la variable j.
  - On pose finalement : $SC(j,i) = 1000 \times \frac{|c(j,i)-\alpha_j|}{\sum_j \alpha_j}$
  - Cette formule n'est valable que dans la situation où tous les coefficients estimés sont positifs et donc min$ic(j,i)$=0, $\forall j$. D'où l'intérêt d'avoir mis les modalités les moins risquées en référence lors de la régression logistique.
- **Contributions des variables** : les variables doivent, dans la mesure du possible, ne pas contribuer à plus de 50 % à la note de score. En effet, dans le cas contraire cela est nuisible au pouvoir prospectif du modèle et il y a un risque de détérioration de la performance dans le cas où la variable deviendrait « instable ».

  - contribution d'échelle CTR_j de la variable j : $CTR_j = \frac{\max_i SC(j,i)}{10}$

  - contribution q_j au score de la variable j : $q_j = \frac{\sqrt{\sum_{i=1}^{m} p_i (SC(j,i)-\overline{SC_j})^2}}{\sum_{k=1}^{n} \sqrt{\sum_{i=1}^{m} p_i (SC(k,i)-\overline{SC_k})^2}}$

  Avec :

  -p_k : part de la population sur la modalité k de la variable j ;

  -$\overline{SC_j}$: note moyenne pondérée (par les effectifs) de la variable j ;

  -m : nombre de modalités de la variable j ;

  -n : nombre de variables retenues dans le modèle.

Comment rendre lisible et interprétable mon modèle ?

Exemple de grille

Variable	Modalité	Répartition	Taux de défaut	Poids	Contribution score	Contribution échelle
Solde moyen sur compte dépôt sur les 3 derniers mois	Inférieur à - 50€ inclus	5.17%		0	9.51%	10.71%
	Entre - 50€ et 250€ inclus	59.82%		65.78		
	Supérieur à 250€	35.02%		107.09		
Encours d'épargne	Inférieur à 80€ inclus	65.27%		0	16.24%	12.93%
	Entre 80€ et 600€ inclus	19.81%		30.57		
	Supérieur à 600€	14.92%		129.25		
		14.98%		0	19.24%	13.85%
		9.83%		38.4		
		11.61%		80.32		
		63.57%		138.5		
		66.49%		0	38.84%	22.86%
		33.51%		228.57		
Catégorie socioprofessionnelle du client	Artisans, Commerçants, Chefs d'entreprise, Ouvriers, NR	8.78%		0	7.90%	11.08%
	Autres	84.12%		51.83		
	Cadres, Professions supérieures	7.10%		110.76		
Interdit chéquier	Clients interdits de chéquier	1.60%		0	5.44%	12.05%
	Clients non interdits de chéquier	98.40%		120.50		
		0.23%		0	2.83%	16.53%
		99.77%		165.32		

Comment rendre opérationnel mon modèle ?

6. Construction d'une échelle de risque (segmentation en CHR)

Travail : A l'aide des valeurs prises par le score, on réalise la segmentation des individus en classes. L'objectif du processus de segmentation en classes est de regrouper les individus, selon les valeurs prises par le score, dans des classes homogènes vis-à-vis du critère d'intérêt.

- **La segmentation en classes doit a minima répondre aux critère suivants :**
  - ✓ Avoir des classes homogènes vis-à-vis du critère d'intérêt ;
  - ✓ Avoir une différenciation appropriée entre les classes (30% d'écart relatif) ;
  - ✓ Avoir un nombre minimal de contreparties par classe (au moins 1% ou 500 individus en défaut).

- **Validation des classes de risque :** s'assurer de leur stabilité en volume et en risque

Classe homogène de risque	Valeur du score	Effectif	Taux de défaut
CHR 1	[746 ; 1000]	31,33%	0,64%
CHR 2	[691 ; 745]	14,41%	1,75%
CHR 3	[641 ; 690]	10,48%	2,69%
CHR 4	[591 ; 640]	12,52%	3,75%
CHR 5	[556 ; 590]	11,25%	5,41%
CHR 6	[451 ; 555]	8,13%	10,57%
CHR 7	[371 ; 370]	4,84%	17,28%
CHR 8	[251 ; 370]	3,97%	25,64%
CHR 9	[146 ; 250]	2,07%	40,93%
CHR 10	[0 ; 145]	0,99%	59,48%

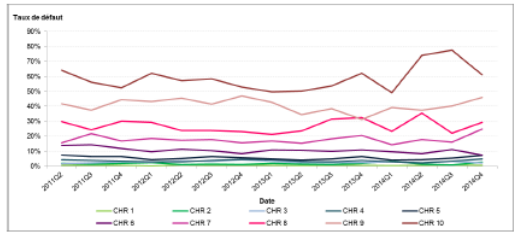


Figure 19: Analyse de la stabilité dans le temps des classes de score pour le modèle à 10 classes

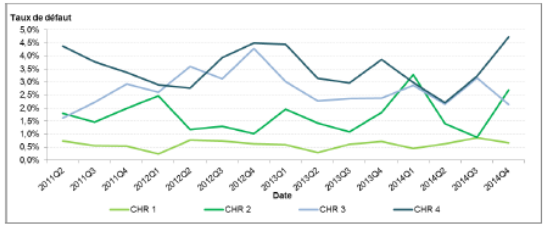


Figure 20: Analyse de la stabilité dans le temps des classes de score dont le taux de défaut est faible, pour le modèle à 10 classes