



12 novembre 2024

# Techniques et méthodes de scoring

*Projet 2024-2025*

*3<sup>e</sup> année – Gestion des Risques*



Ma vie. Ma ville. Ma banque.



Ma vie. Ma ville. Ma banque.

# Introduction

## Informations pratiques liées au Projet de Techniques et Méthodes de Scoring

- Présentation de l'intervenant ;
  - > Mail : [clement.gabas@lcl.fr](mailto:clement.gabas@lcl.fr)
  
- 3 x 6h de TD :
  - > 12/11/24 : de 09h45 à 12h45 et de 14h à 17h ;
  - > 25/11/24 : de 09h45 à 12h45 et de 14h à 17h ;
  - > 12/12/24 : de 09h45 à 12h45 et de 14h à 17h ;
  
- 3h de présentation orale intermédiaire de l'avancée des travaux entre les séances 2 et 3
  - > 05/12/24 : de 9h45 à 13h à distance



Ma vie. Ma ville. Ma banque.

# Introduction

## Objectifs pédagogiques du Projet de Techniques et Méthodes de Scoring

### Objectif pédagogiques du cours

Les objectifs pédagogiques du cours, hautement importants pour votre parcours professionnel, en risque de crédit, de marché, ou ailleurs, sont :

- Connaître et maîtriser la méthodologie de construction d'un **score statistique** et des notions sous-jacentes, notamment :
  - La définition du périmètre d'étude ;
  - La construction de la base d'analyse ;
  - La sélection des variables ;
  - L'estimation du modèle *standard* et l'analyse des performances ;
  - L'estimation de modèles concurrents et l'analyse de leurs performances ;
  - La restitution explicable de chaque type de modèle.
- Etre capable de restituer la technicité et les résultats du score dans un document technique écrit ;
- Etre capable de restituer au bon niveau de technicité les résultats devant les métiers de LCL (équipe de Modélisation des Risques et Directeur des Risques).



Ma vie. Ma ville. Ma banque.

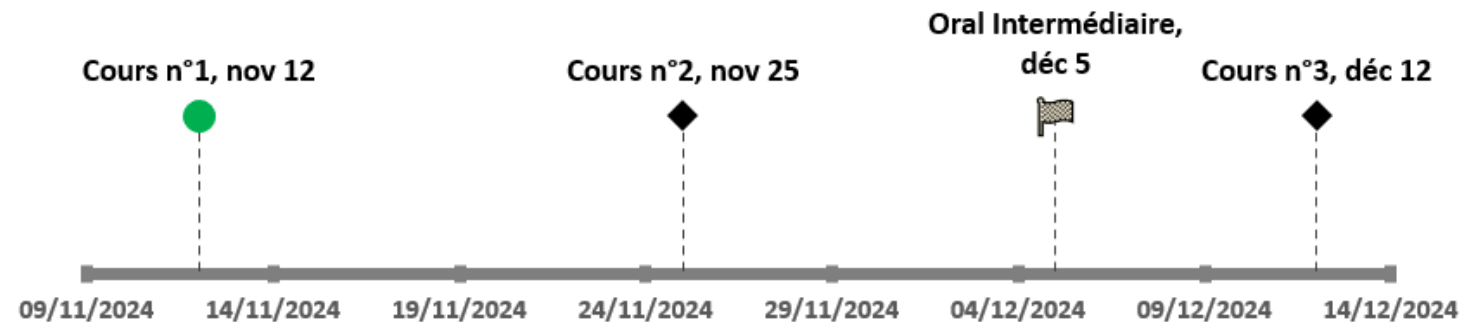
# Introduction

## Attendus pédagogiques du Projet de Techniques et Méthodes de Scoring

### Déroulement des séances et attendus pédagogiques du cours

Les séances se dérouleront de la façon suivante :

- Cours 1 (9h45-11h10) : présentation du contexte et des objectifs ;
- Cours 1 (11h20-13h) : présentation de la méthodologie de construction d'un score ;
- Cours 1 (14h-17h) : prise en main des données et travaux sur le score ;
- Cours 2 : travaux sur le score ;
- Oral intermédiaire : oral de présentation des travaux, comptant pour la note finale ;
- Cours 3 : travaux sur le score et notation *type kaggle* de vos scores ;
- Fin Janvier 2025 : rendu définitif.





Ma vie. Ma ville. Ma banque.

## 1 Présentation de LCL - Modélisation

## 2 Méthodologie de construction d'un score

## 3 Présentation du projet

## 4 Scores concurrents





Ma vie. Ma ville. Ma banque.

# Présentation générale de LCL

## LCL et le Groupe Crédit Agricole





# Présentation générale de LCL

## LCL et le Groupe Crédit Agricole

Ma vie. Ma ville. Ma banque.

### Banque de proximité en France



### Banque de financement et d'investissement



### Crédit à la consommation



### Gestion d'actifs



### Banque de proximité à l'international



### Immobilier



### Assurances



### Banque Privée



### Activités spécialisées



### Crédit bail & affacturation



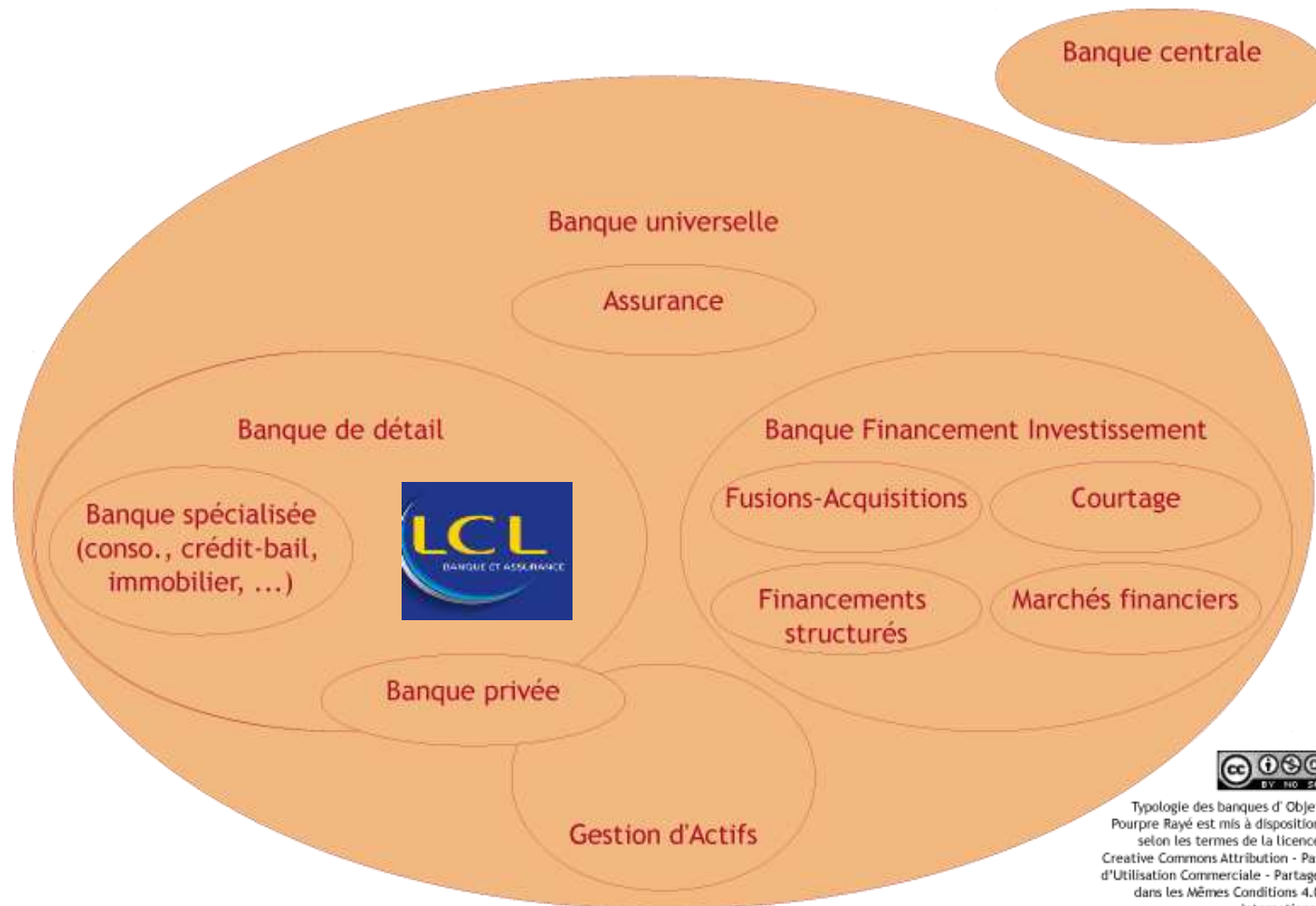




Ma vie. Ma ville. Ma banque.

# Présentation générale de LCL

## Typologie des banques

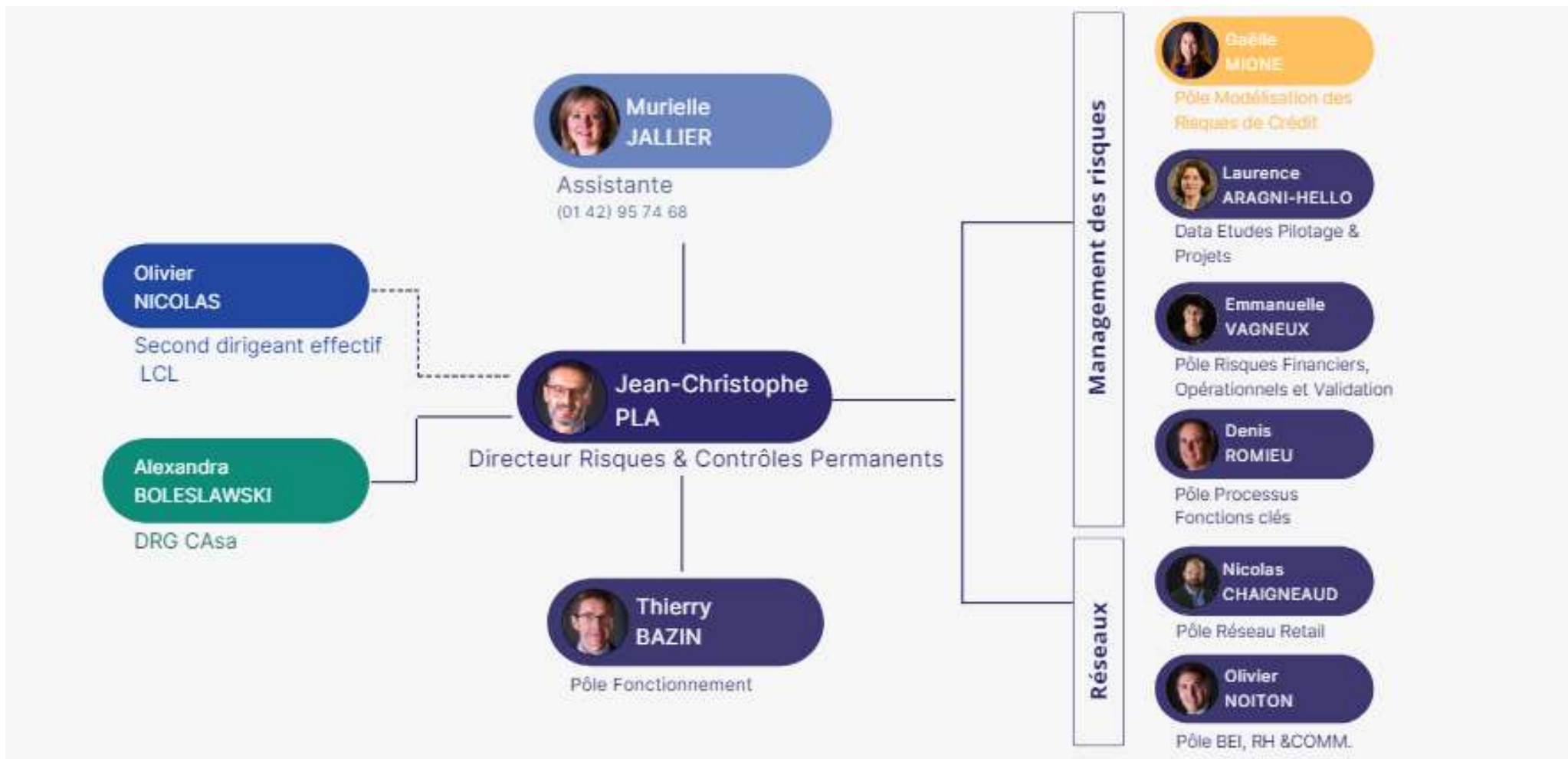


Typologie des banques d'Objet  
Pourpre Rayé est mis à disposition  
selon les termes de la licence  
Creative Commons Attribution - Pas  
d'Utilisation Commerciale - Partage  
dans les Mêmes Conditions 4.0  
International



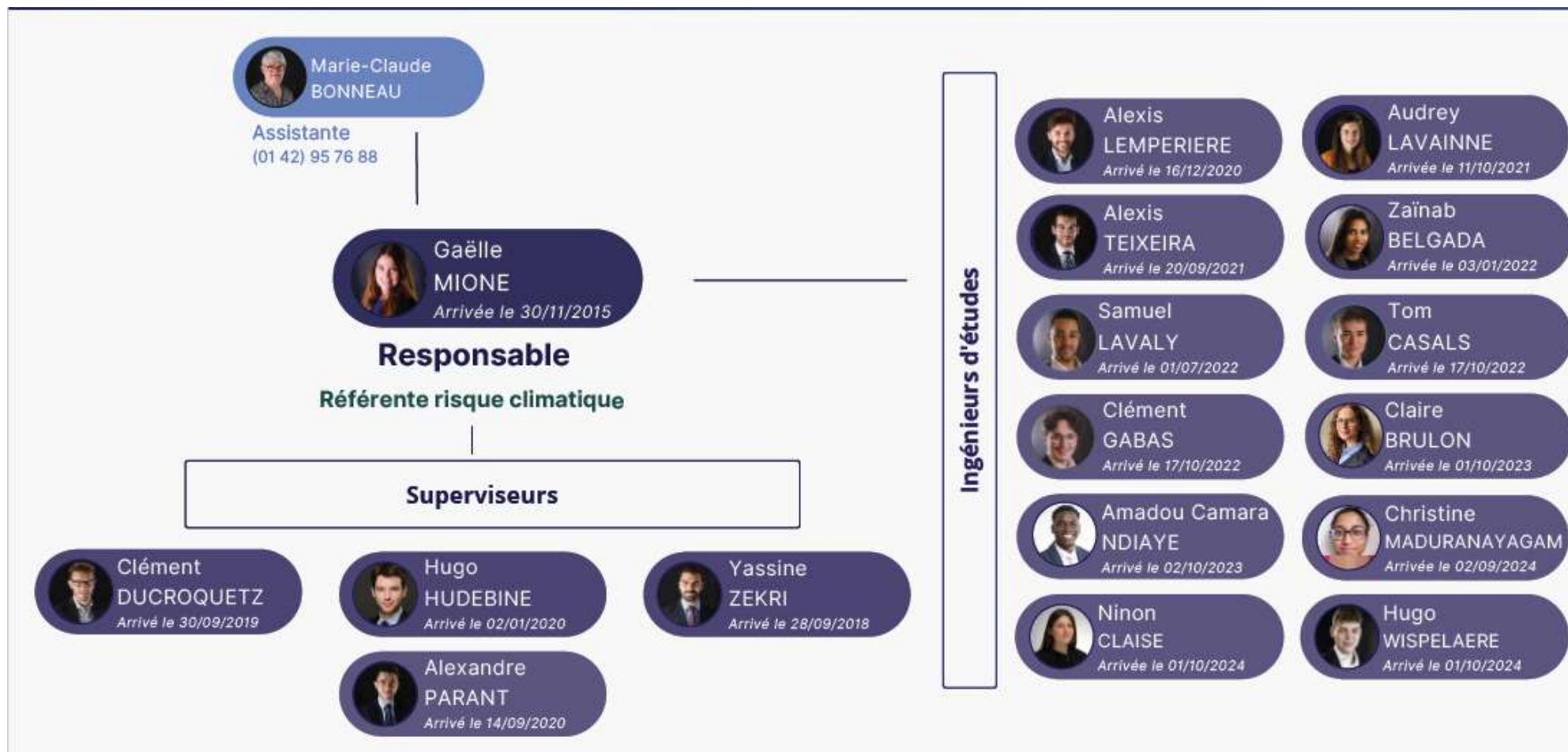
# Présentation équipe RCP/Modélisation

## Organigramme Direction des Risques et Contrôles Permanents



# Présentation équipe RCP/Modélisation

## Organigramme RCP\Modélisation



# 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

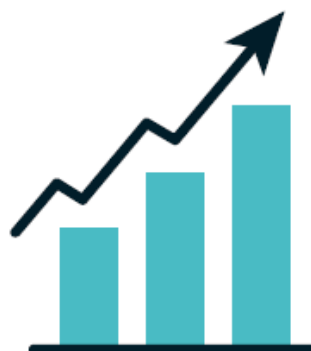
Construction, suivi et mise à jour des scores/outils d'aide à la décision utilisés par le Réseau (e.g. scores immobiliers, score d'octroi et pré-attribution aux professionnels, anticipation des risques, Aide à la Décision du Jour)



### 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

## RCP Modélisation



### 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



Ma vie. Ma ville. Ma banque.

## 1 Présentation de LCL - Modélisation

## 2 Méthodologie de construction d'un score

## 3 Présentation du projet

## 4 Scores concurrents





# Les étapes de construction d'un score

## Principe d'un modèle statistique

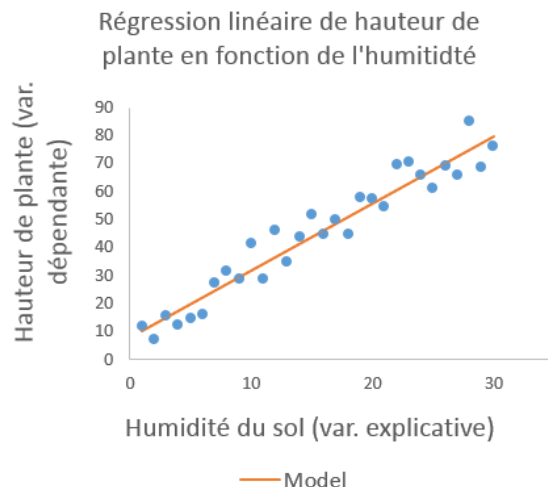
### Qu'est-ce qu'un modèle statistique ?

Un **modèle statistique** est une description mathématique approximative d'un mécanisme observable. A travers l'observation de réalisation de ce mécanisme, que l'on suppose être un processus stochastique et non un processus déterministe, un modèle a pour objectif de répliquer au mieux le mécanisme observé.

#### Un modèle :

- **est une représentation idéalisée de la réalité** - en effet, un modèle ne prend en considération qu'un nombre fini de causes pour expliquer les conséquences observées. On parle alors des « paramètres » du modèles (variables en entrées, hyperparamètres, etc...);
- **fait des hypothèses explicites sur les processus étudiés** (loi de distribution, famille de paramétrisation, modèle linéaires, mixtes, structurés, ...);
- **ces hypothèses peuvent être fausses ou approximatives.**

Exemple classique : modèle expliquant la hauteur d'une plante via l'humidité du sol, par une relation linéaire.



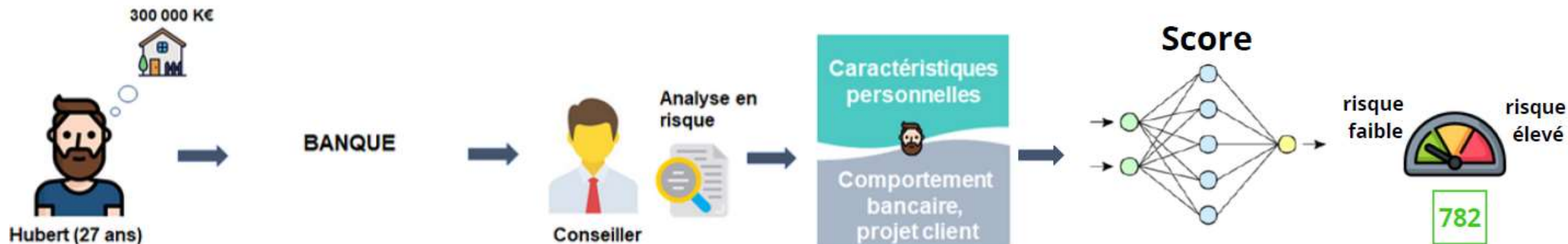
On parle alors :

- de **variable à expliquer** (taille de la plante – peut être numérique ou catégorielle) ;
- des **variables explicatives** (humidité du sol – peuvent être numériques ou catégorielles) ;
- du **type de modèle** (régression/classification selon le type de la variable à expliquer, linéaire, catégorielle, forêt, gradient, réseau de neurones, etc... selon l'approche et les hypothèses retenues.)

Pour plus de détails (identification, paramétrisation, test des hypothèses, mesure des performances, ...), je vous invite à reprendre vos cours de 1A et 2A. En particulier pour ce projet, le cours de GLM, de tests statistiques et de modèles statistiques plus avancés.

# Les étapes de construction d'un score

## Principe d'un score



- Age
- Catégorie socio-professionnelle
- ...
- Nb. de lignes créditrices
- Nb. de jours débiteurs
- ...

52

80

31

109

78

...



Un score est un modèle statistique qui permet, à partir d'un système de points, de donner la probabilité de réalisation d'un événement binaire.



Sur la base de l'analyse de données de tout types (signalétique, bancaire, Marketing, risque), le score permet de regrouper les individus notés avec des caractéristiques similaires en classe (appelées Classes Homogènes de Risques – CHR).



Le score est utilisé dans le processus de décision, souvent en accord avec d'autres systèmes ou règles expertes.



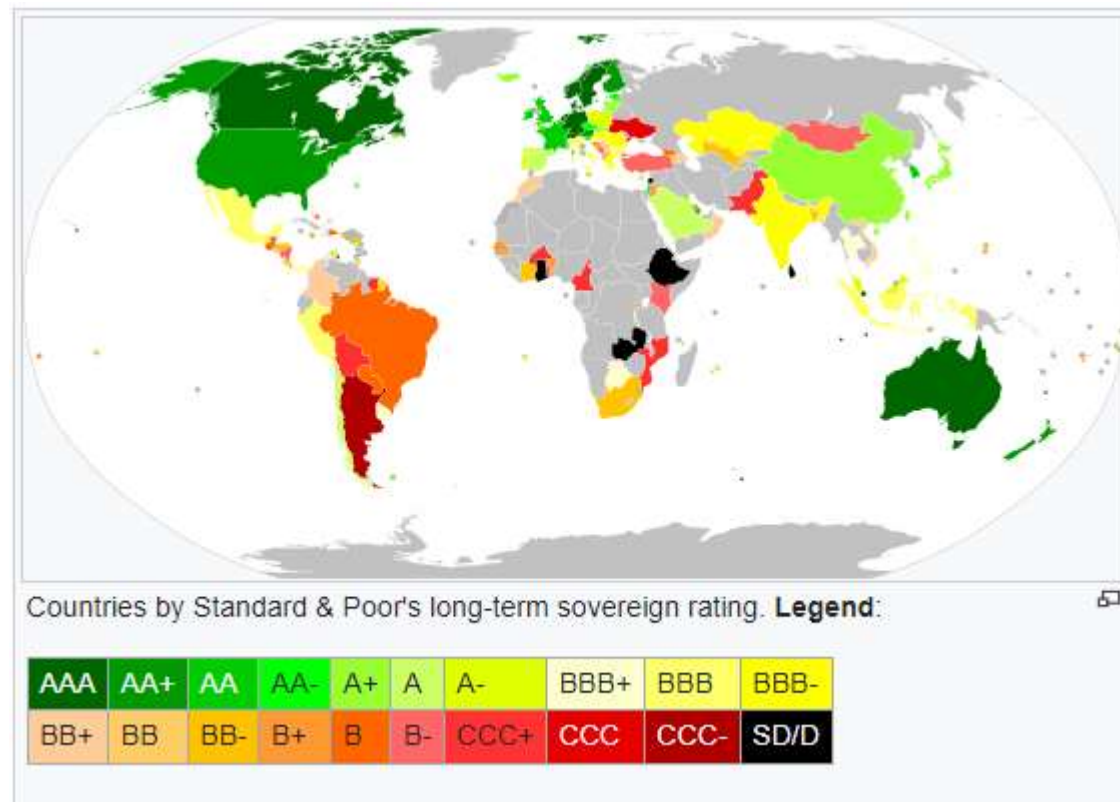
# Les étapes de construction d'un score

## Principe d'un score

### Que cherche-t-on à modéliser dans un score

A travers l'utilisation d'un score, on cherche à modéliser un évènement binaire. On liste ci-dessous les principaux cas d'usage d'un score dans le domaine bancaire (liste non exhaustive) :

- **Notation du risque des clients**, à l'octroi comme durant la vie du contrat :
  - *Retail* : utile pour l'accord ou non du crédit, la délégation et les potentielles renégociations au cours du temps ;
  - *Corporate* : utile pour l'accord et le pricing du crédit ainsi que le tarif des potentielles renégociations ;
  - *Souverain* : utile pour le pricing de l'obligation d'état ;
  - Exemple concret : les notes des agences de notations (Moody's, S&P, ...)
- **Détection d'anomalie** : un score peut être utilisé pour détecter des anomalies, telles que :
  - Détection de fraudes (chèques, documents, virement, ...) ;
  - Blanchiment d'argent, détournement de fonds, ...
  - (dans d'autres domaines) détection de cancer, détection de malfaçon industrielle, ...
- **Ciblage client**, pour des opérations diverses (ciblage d'interventions financières ou de vente de produit, ciblage promotionnel, Marketing, attrition, ...)
- **Variations de marché** : un score peut être utilisé pour prédire les variations à la hausse ou la baisse d'un actif sur le marché, en modélisant par exemple la probabilité que l'actif monte selon divers critères (conjecture, AR, etc...)



# Les étapes de construction d'un score

## 1. Définition des objectifs et périmètre d'étude



### Quelles questions se poser ?

Avant de commencer à travailler sur un score, la première étape importante consiste à définir les objectifs recherchés et la finalité du projet.

#### Théorie

##### ➤ Intérêts du score, résultats attendus et population cible

- Nouveau score ou refonte ?
- Problématique
- Type de produit concerné et unité statistique (client, contrat, ...)
- Résultats et gains attendus ?
- Utilisation opérationnelle

##### ➤ Données de modélisation

- Critère cible à modéliser (défaut de paiement, impayés, fraude, hausse des prix ...)
- Historique de données suffisamment long, stable, fiable et robuste.

##### ➤ Segmentation

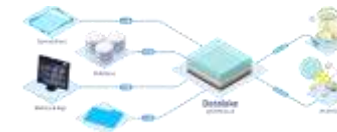
- Dans de nombreux cas, mettre en place des méthodes de notation différentes sur des typologies de clients différents s'avère justifié, notamment lorsque les comportements entre les sous populations sont hétérogènes.
- Exemple concret : client connu vs clients récent vs prospect bancaire.

#### Pratique

- Le score attendu sera présenté dans les slides suivantes. Ce travail a été réalisé en amont par l'équipe de modélisation des risques de LCL.

# Les étapes de construction d'un score

## 2. Inventaire et collecte des données



### Les données de travail

Plus de 80% des résultats d'un score proviennent de la qualité des données utilisées. Sans des données de qualité, même le meilleur ingénieur ne pourra rien faire. Il faut donc largement s'attarder sur la qualité des données.

#### Théorie

##### ➤ Cartographie des données

- Données internes
- Données externes (open source)
- 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 pour construire la *base d'analyse*
- Clé de jointure
- Temps de traitement

#### Pratique

##### ➤ Cartographie des données

- Prise en main du dictionnaire de données et identification de la variable à expliquer et des variables candidates au modèle.
- Identifier les données quantitatives et qualitatives de la base

##### ➤ Type de données

- Le score attendu sera présenté dans les slides suivantes. Ce travail la été réalisé en amont par l'équipe de modélisation des risques de LCL.

##### ➤ Collecte des données

- Le score attendu sera présenté dans les slides suivantes. Ce travail la été réalisé en amont par l'équipe de modélisation des risques de LCL.

# Les étapes de construction d'un score

## 3. Construction de la base d'analyse



### La base d'analyse (1/5)

Une fois la base d'analyse construite, documentée et fiabilisée, le travail le plus important de la modélisation peut commencer :

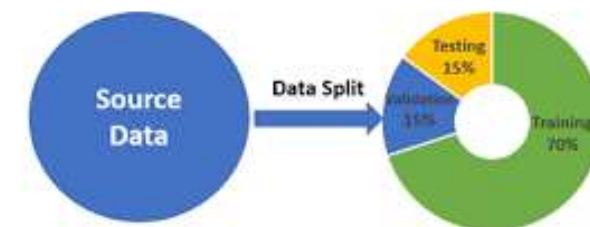
#### Théorie

#### ➤ Construction des échantillons (1/2) :

- La base d'analyse étant bien souvent très volumineuse, un premier échantillonnage peut être réalisé pour réduire la volumétrie. Auquel cas, il convient de s'assurer que les données de la base réduite restent représentatives de la base d'analyse initiale.
- Par ailleurs, **la mesure de la prédictivité des scores sur l'échantillon sur lequel a été effectuée la modélisation conduit à des résultats biaisés vers des performances optimistes** (on parle de **surapprentissage**). *Cela repose sur le fait que l'estimation des paramètres à l'origine du calcul des scores est basée sur un procédé d'optimisation (par exemple la recherche de la vraisemblance maximum dans le cas d'une régression logistique), ce qui implique que le modèle cherche par définition les valeurs des paramètres fournissant la meilleure adéquation entre le critère à modéliser prédit et celui observé.*
- On travaille alors avec 3 échantillons :
  - **L'échantillon d'apprentissage** (*training set* - souvent 70/80% de la base d'analyse). *L'échantillon d'apprentissage sert au développement et à l'estimation du modèle. C'est sur cet échantillon que les variables vont être construites et les coefficients estimés ;*
  - **L'échantillon de test** (*testing set* - les 20/30% restant). *L'échantillon test permet de faire une validation structurelle en vérifiant le modèle sur une population similaire (même période d'observation) mais qui n'a pas servi à la construction du modèle ;*
  - Un potentiel **échantillon hors temps** (*validation set*), qui permettra de valider le drift des résultats sur des données indépendantes temporellement des données d'apprentissage. *L'échantillon Out of Time (OoT) permet de faire une validation, appelée validation hors du temps, sur une période différente de celle de la construction du modèle. La performance du score doit être vérifiée sur l'échantillon OoT.*

On parle alors **d'échantillonnage des données (data splitting)**.

- Ainsi, c'est en comparant les performances obtenues sur les deux sous-échantillons qu'on jugera de la robustesse du modèle (des performances semblables permettront de conclure à un modèle robuste).



# Les étapes de construction d'un score

## 3. Construction de la base d'analyse



### La base d'analyse (2/5)

#### ➤ Construction des échantillons (2/2) :

##### ➤ Points d'attentions :

- Pour assurer la représentativité de ces échantillons sur le futur périmètre d'application on procède généralement à un **tirage aléatoire sans remise, stratifié** par le critère à modéliser.
- Il est souhaitable d'avoir des **observations indépendantes** (et donc de limiter au maximum par exemple les observations sur un même client à des dates consécutives).

##### ➤ Analyse de la représentativité des échantillons : le découpage de la base en plusieurs échantillons nécessite une analyse de la représentativité de ces nouveaux échantillons par rapport à la base initiale. On vérifie notamment (liste non exhaustive) :

- Que la distribution temporelle du critère cible ne varie pas trop entre les échantillons (seuil arbitraire selon les données) ;
- Que la volumétrie reste stable au cours du temps.

# Les étapes de construction d'un score

## 3. Construction de la base d'analyse



### La base d'analyse (3/5)

Une fois la base d'analyse construite, documentée et fiabilisée, puis splittée en train, test et oot, une pré-sélection des variables candidates à l'explication de la variable cible est réalisée :

- **Pré sélection des variables** : on ne conserve que les données
  - **Fiables** (peu de NA ou de valeurs aberrantes, ou alors ces données sont explicables) ;
  - **Disponibles** (données disponibles de façon homogènes dans le temps – par exemple l'apparition/la disparition d'une modalité au cours du temps, le changement des règles de gestion, de branchement IT qui pourraient dérégler la source, ...) ;
  - **Pertinentes** (on ne perd pas de temps sur des données non pertinentes) ;
  - **Légales** : les réglementations multiples sur la protection des consommateurs et l'utilisation de modèles de ML et d'IA dans l'industrie encadrent la légalité de l'utilisation de certaines données. En France, la CNIL (avec la RGPD) et en Europe l'AI Act (COM/2021/206) encadrent en outre le type de données conformes à l'utilisation. En particulier, des données comme l'orientation sexuelle, le genre, ou même la nationalité ou le statut marital sont très encadrées et ne peuvent être utilisées comme on le souhaite.
  - **ATTENTION** : on ne supprime pas une observation car elle a des NA! Car dans la pratique, ces observations devront être notées par le score, et ceux malgré la potentielle absence d'une donnée. On verra dans la suite qu'il convient plutôt de catégoriser les NA dans des catégories spécifiques! De façon générale, on ne supprime aucune observation de nos bases. Si la donnée est observée, c'est qu'elle existe et donc qu'il faut l'utiliser pour être modélisée.



# Les étapes de construction d'un score

## 3. Construction de la base d'analyse



### La base d'analyse (4/5)

Une fois la base d'analyse construite, documentée et fiabilisée, puis splittée en train, test et oot et les variables candidates présélectionnées, on réalise un dernier travail sur la base d'analyse :

- **Création d'indicateurs**: cette opération vise à créer des indicateurs plus pertinents, d'un point de vue métier, à partir des données brutes de la base d'analyse. Il peut d'agir :
  - de remplacer des dates (par des durées, des anciennetés, des âges, ...) ;
  - de calculer des ratios ;
  - de synthétiser une évolution temporelle (créer des agrégats, de données mensuelles, semestrielles, calculer des évolutions sur les 1/3/6 derniers mois, ...) ;
  - de modifier les unités de mesure (par exemple passer d'une mesure en centimes d'Euros à une mesure en Euros) ;
  - Ou d'effectuer n'importe quelle autre transformation, linéaire ou non, d'une donnée qui la rendrait plus apte à expliquer la donnée cible (passage au log, logit, exponentielle, valeur absolue, rendements, ...)

# Les étapes de construction d'un score

## 3. Construction de la base d'analyse

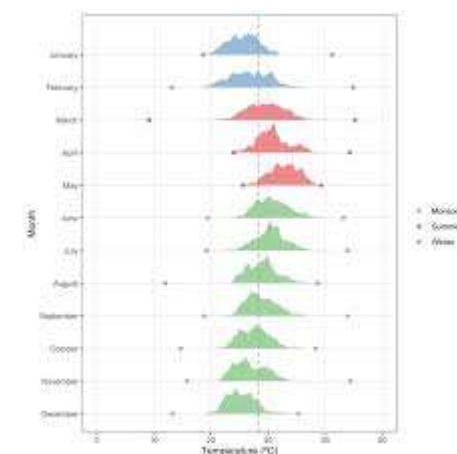
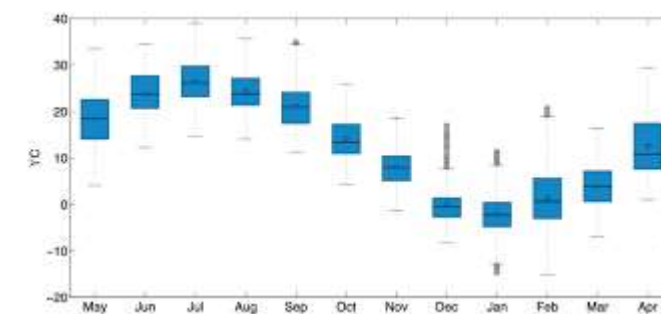


### La base d'analyse (5/5)

#### Pratique

#### ➤ Construction des échantillons

- La base fournie n'est pas très volumineuse. Une première réduction de sa taille n'est pas nécessaire.
- Réaliser le data splitting
  - Implémentation :
    - **Python** : `sklearn.model_selection.train_test_split`
    - **SAS** : PROC SURVEYSELECT
- Analyser la représentativité des échantillons
  - Par des représentations graphiques (évolutions de la distribution mois par mois via boxplot, densité, ...)
  - Par des chiffrages (taux de variations mensuels d'une donnée, fréquence d'une modalité, ...)
  - Détecter les doublons.
- L'analyse de conformité des données a été réalisée en amont par l'équipe Modélisation de LCL
- Créer des indicateurs





Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## 4. Sélection des variables



### STOP POINT

**Pour toute la suite des slides, et sauf mention contraire, les études sont à réaliser sur l'échantillon d'apprentissage uniquement!**

# Les étapes de construction d'un score

## 4. Sélection des variables



### Sélection des variables (1/3)

Une analyse exploratoire est une étape indispensable une fois l'échantillon d'apprentissage constitué afin de détecter quelles sont les variables qui vont le mieux expliquer le critère à modéliser et quel est le lien entre elles. A l'issue de cette étape, nous obtiendrons une liste de variables prêtes à être testées dans le modèle.

Les variables sont séparées en deux groupes : les données catégorielles (ou qualitatives) et les données quantitatives.

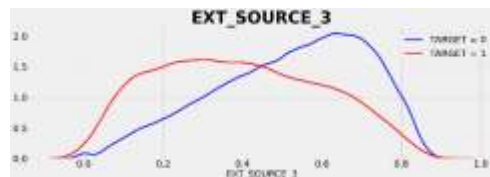
#### Statistiques descriptives univariées

On commence par décrire chaque variable candidate sur l'échantillon d'apprentissage :

##### Variables quantitatives

- On étudie la distribution de la variable (tendance, dispersion, quantiles, ... ) ;
- On représente graphiquement la distribution (densité, boxplot, histogramme).

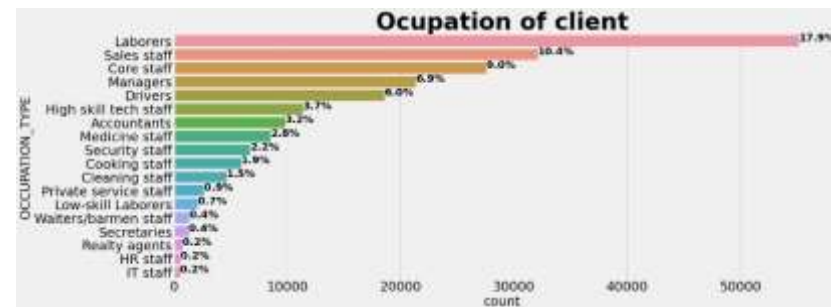
- En python : fonction de stats numpy, fonction graphique plotly ou seaborn ;
- En SAS : proc means et proc sgplot



##### Variables qualitatives

- On étudie la fréquence de chaque modalité ;
- On représente graphiquement des barplot empilés ou des camemberts.

- En python : fonction de stats numpy et d'agrégation pandas .agg() ;
- En SAS : proc freq et proc sgplot



# Les étapes de construction d'un score

## 4. Sélection des variables



### Sélection des variables (2/3)

Une fois que chaque variable candidate a été résumée, il est nécessaire de s'intéresser aux liaisons qu'il y a entre elles. L'objectif final est d'avoir :

- **Un lien fort entre la variable à expliquer et les variables explicatives** : l'objectif du score est de trouver les variables qui permettront d'expliquer au mieux le critère à modéliser. Les variables qui sont testées dans le modèle doivent donc avoir un lien suffisamment fort avec la variable à expliquer. Plus la variable explicative est corrélée avec la variable à expliquer, plus cette variable sera en mesure de discriminer le critère à modéliser ;
- **Un lien faible entre les variables explicatives** : une des hypothèses primordiales à respecter dans le cadre de la construction d'un score (et dans la plupart des modèles statistiques en général) est d'éviter la multicolinéarité entre les variables explicatives. Un lien trop fort entre les variables explicatives pourrait entraîner de la multicolinéarité. Les facteurs doivent donc être les plus décorrélés possible.

### Liaisons entre les variables

Divers inducteurs statistiques existent pour déterminer l'intensité de la liaison entre deux variables selon leur type :

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

# Les étapes de construction d'un score

## 4. Sélection des variables



### Sélection des variables (3/3)

#### Liaisons entre les variables

On garde les données les plus liées avec la cible en ordonnant les données par :

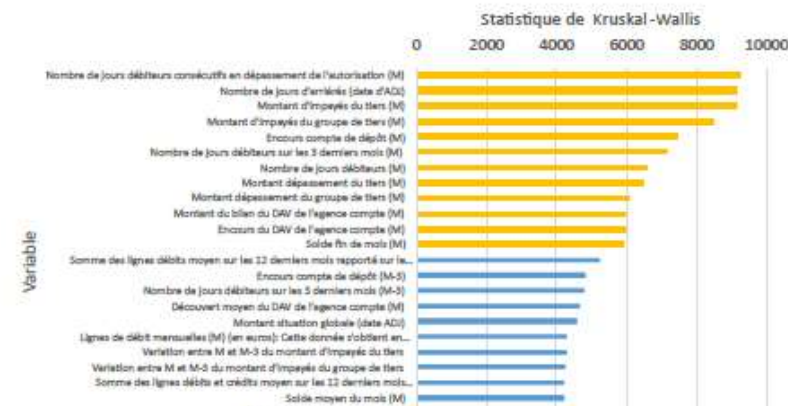
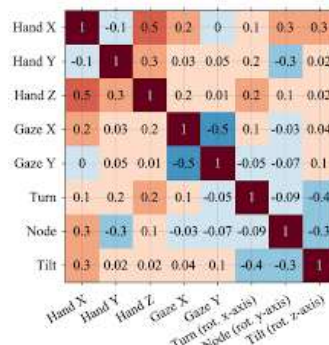
- V de Cramer avec la cible pour les variables qualitatives ;
- Statistique de KW avec la cible pour les variables quantitatives ;

	Python	SAS
<b>Pearson</b>	scipy.stats.Pearsonr	proc corr
<b>Spearman</b>	scipy.stats.spearmanr	proc corr
<b>Kruskal Wallis</b>	scipy.stats.kruskal	proc npar1way
<b>Cramer</b>	à coder à partir de scipy.stats.chi2_contingency	proc freq

On crée alors une matrice de corrélation de toutes les variables quantitatives entre elles (et avec la cible) et une matrice de toutes les variables qualitatives entre elles (et avec la cible), ordonnées dans l'ordre décroissant (de la variable la plus liée à la cible à la variable la moins liée).

On passe en revue les variables une à une dans l'ordre.

- Si la variable est liée à la cible et peu corrélée avec une autre donnée davantage liée à la cible : on la garde ;
- Si la variable est peu liée à la cible OU qu'elle est liée à la cible mais corrélée avec une autre donnée davantage liée à la cible : on la drop.



On essaye ensuite de croiser les variables dropées entre elles pour créer de nouvelles informations intéressantes et on recalcule la matrice et les règles de sélection pour voir si on conserve la nouvelle variable créée ou pas.

Exemple : si les variables d'ancienneté de la relation client et d'âge de l'entreprise ne passent pas le test, on peut les fusionner pour créer une variable « l'entreprise a-t-elle été bancarisée ailleurs avant? ».



# Les étapes de construction d'un score

## 5. Construction de modalités



### Discrétisation des variables quantitatives

Lors de l'élaboration d'un modèle de score, il est d'usage d'effectuer un **découpage des variables quantitatives** en classes (discrétisation), voire en indicatrices. Cette opération présente en effet de nombreux avantages :

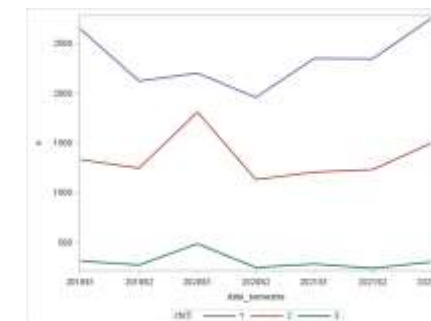
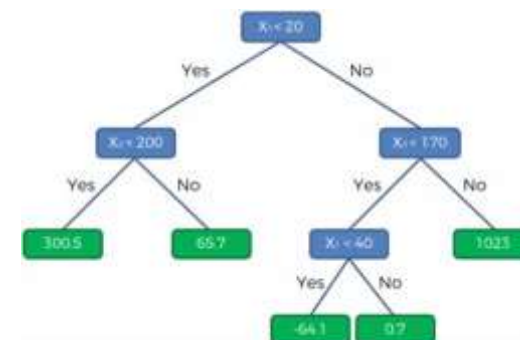
- ✓ Meilleure lisibilité et interprétation « métier » de la grille de score (recoder les modalités et variables avec des noms trop longs ou imprécis) ;
- ✓ Gain d'information : malgré un paradoxe apparent, on gagne généralement en précision en découpant en classes une variable quantitative (car les liaisons purement monotone entre variable quantitative et cible sont très rares) ;
- ✓ Plus grande robustesse d'un modèle lorsqu'on modélise un phénomène rare ;
- ✓ Gestion des valeurs manquantes (regroupées dans une classe spécifique) et neutralisation des valeurs extrêmes ;
- ✓ Prise en compte d'effets non linéaires

La discrétisation doit créer des classes à l'intérieur desquelles **le phénomène à prédire est relativement stable, en volume et en risque**, au cours du temps. Par ailleurs, **le risque entre les classes doit être le plus différencié possible**.

La discrétisation peut s'effectuer par diverses approches :

- Créations automatiques de classes qui maximisent/minimisent un indicateur (chi<sup>2</sup> normé par exemple) ;
- Découpage de la donnée en centiles et regroupement des centiles adjacents ayant des taux de variable cible similaires ;
- Discrétisation supervisée par arbre de décision, en injectant si nécessaire une matricule de coût.

Il convient ensuite de s'assurer que les bornes des modalités sont pertinentes et interprétables d'un point de vue métier. Des rééquilibrages « à la main » doivent donc avoir lieu (exemple : borne optimale à 4€94 sur le solde de fin de mois, arrondi à 5€ dans la grille / borne optimale de 18ans et 14 jours, arrondi à 18 ans, etc...).



# Les étapes de construction d'un score

## 5. Construction de modalités



### Regroupement des modalités des variables qualitatives

La réduction du nombre de modalités consiste à :

- Regrouper les modalités qui sont trop nombreuses ou dont les effectifs sont trop petits ;
- Regrouper les modalités qui ont la même signification métier et un taux de critère équivalent (comportement similaire vis-à-vis du critère à modéliser).

Le regroupement de modalités d'une variable qualitative répond donc aux mêmes contraintes et objectifs que la discrétisation des variables quantitatives, notamment de **stabilité en volume et en risque**.

**Points d'attention** : la consultation des métiers lors de cette étape est essentielle. En effet, il est possible qu'ils aient une idée au préalable des regroupements à faire d'après leur expérience. Ils doivent aussi intervenir afin de valider les regroupements qui auront été réalisés.

Un effet d'interaction (croisement) correspond à une nouvelle variable issue du croisement des modalités de deux (ou plus) variables qualitatives. Les croisements doivent être justifiés par une approche métier pour ne pas croiser deux donner qui n'ont aucun sens métier entre elles (principe des torchons et des serviettes) et doit être interprétable.

# Les étapes de construction d'un score

## 5. Construction de modalités



### Stabilité des variables découpées

Afin de s'assurer que les modalités des variables sont stables et robustes, il faut vérifier certains critères :

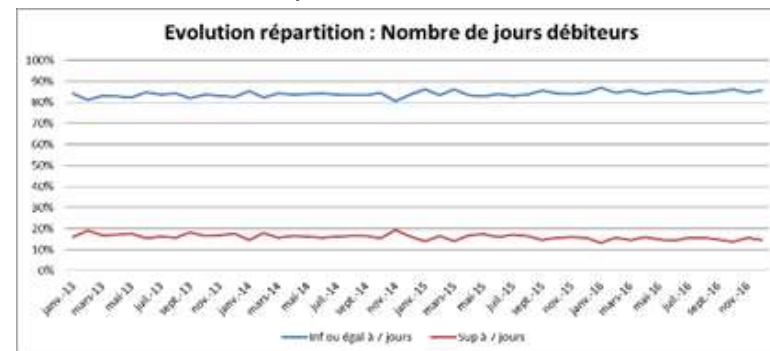
- le taux du critère à modéliser entre les modalités d'une variable doit être croissant (la modalité 1 à le plus faible taux, la nième à le plus fort taux) ;
- Les modalités doivent être stables en taux du critère à modéliser (pas de croisement entre les modalités au cours du temps) et en volumétrie ;
- Un **effectif minimum de 5%** de la population doit être trouvé dans chaque modalité (à l'exception, de variable particulièrement explicative du critère à modéliser et particulièrement fiable, pour exemple, dans le cadre de score de risque, pour des variables de dépassement d'autorisation ou d'impayé un effectif moindre est toléré) ;
- Les modalités doivent être séparée d'au moins **30% relatif** en termes de taux de variable cible.

Si l'un de ces critères n'est pas respecté, un regroupement se fait entre les modalités ayant les taux les plus proches, et les critères sont à nouveau testés.

**Ces critères doivent être vérifiés sur les échantillons de train, de test et OOT concomitamment.**

Une fois toutes les données discrétisées et les modalités regroupées, il faudra recalculer les indicateurs de discrimination utilisés afin de mesurer l'ordre d'importance des variables candidates découpées et le lien avec la donnée cible.

On ne gardera que les données liées à la cible, étant liées à moins de 60% (V de Cramer) avec d'autres données plus liées à la cible.





Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## 6. Construction du modèle

### STOP POINT

Toutes les étapes précédentes, et en particulier la sélection des variables et le découpages en modalité, sont à réaliser quelque soit le type de modèle utilisé dans construction du score.

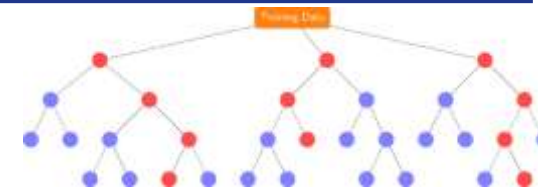
Dans certains cas, pour des modèles plus complexes que des simples régressions, on pourra accepter de ne pas découper certaines variables quantitatives car les modèles à interactions plus complexes gèrent mieux la non monotonie stricte que les modèles linéaires.

Les étapes mentionnées ci avant sont donc primordiales car communes à tous vos modèles, et elles décident à elles seules plus de 85% de la qualité prédictive du modèle final.

**Dans le cadre de ce projet, on attendra de vous d'avoir avancé ces étapes de votre coté entre les séances 1 et 2, afin que l'encadrant puisse vous donner des conseils pour améliorer vos données au cours de la séance 2, ce qui vous permettra d'avoir le temps de peaufiner le score logistique pour l'oral intermédiaire après la séance 2.**

# Les étapes de construction d'un score

## 6. Construction du modèle



### Choix du modèle explicatif

Une fois les données candidates à l'explication de la donnée cible découpées en modalités fiables, il convient d'apprendre le modèle de prédiction de la donnée cible.

On est confronté à un choix : application d'un modèle de classification ou application d'un modèle de régression ?

On a envie de choisir un **modèle de classification**, puisqu'on essaye de prédire l'apparition de la donnée cible binaire (oui ou non).

Nous optons ainsi pour des modèles de **régression sur la probabilité** d'appartenir à chaque classe. Il s'agit d'expliquer au mieux une variable binaire (la présence ou l'absence d'une caractéristique donnée) par des observations réelles.

Parmi ces modèles, on a le choix entre divers types : **régression logistique**, **arbres de décisions (dont RF et variantes)**, **xgboost**, autres algorithmes de Deep Learning, ...

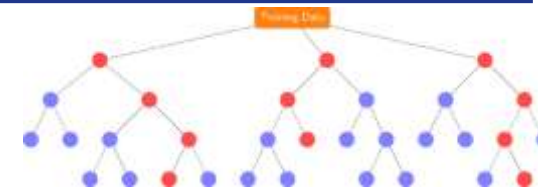
Cependant, en gestion des risques bancaires (dont marché, assurantiels et financiers), les modèles sont audités par la BCE et l'EBA et sont régis par des réglementations très précises sur ce qui est autorisé ou non, ce qui est encouragé ou non, etc..

En particulier, il est obligatoire pour un modèle réglementaire d'être :

- **Interprétable** (on peut comprendre comment le modèle fonctionne) ;
- **Explicable** (on peut expliquer de façon exacte la prédiction réalisée par le modèle pour chaque individu, accessible à un néophyte) ;
- **Implémentable** (c'est-à-dire pouvant être implémenté dans les chaînes IT d'une banque ou d'un organisme financier. Ces chaînes IT sont souvent très anciennes, dans des langages peu souples, notamment pour des raisons de cybersécurité, mais aussi pour des raisons historiques) ;
- (mais aussi auditable, performant, robuste et stable)

# Les étapes de construction d'un score

## 6. Construction du modèle



### Choix du modèle explicatif

Ainsi, la régression logistique garde les faveurs de l'industrie car :

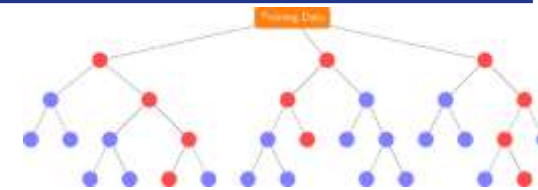
- Les performances sont sensiblement aussi bonnes, voir meilleures, que les modèles plus complexes ;
- Le modèle est parfaitement interprétable, explicable et implémentable ;
- Il est très aisé d'obtenir une grille de lecture des résultats (appelée grille de score).

Dans ce projet, on réalisera un score par **régression logistique**, puis un autre score par un **autre modèle de votre choix**, et on comparera les résultats et l'explicabilité des résultats.



# Les étapes de construction d'un score

## 6. Construction du modèle



### Rappels sur la régression logistique

Lorsqu'on souhaite expliquer une variable discrète  $Y \sim \text{Bernoulli}(p)$ , avec  $p = P(Y=1)$ , on s'intéresse à l'influence des covariables  $X$  (variables explicatives découpées en modalités) sur la paramètre  $p$ . On cherche donc  $p(x)$  avec  $x$  le vecteur de covariables.

Pour estimer les valeurs de  $p(x)$  (qui dépendent donc de  $x$ ), on pense naturellement à utiliser une régression linéaire, la cible  $p(x)$  étant une donnée continue. Cependant, un modèle linéaire de la forme  $p(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$  pouvant prendre des valeurs négatives et/ou supérieures à 1 en inadéquation avec les propriétés d'une probabilité, on utilise une fonction de lien que l'on note  $G$ .

Cette fonction de lien, qui n'est rien d'autre qu'un changement de variable bien choisi, permet de pouvoir mener l'étude suivante :  $G(p(x)) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$ .

Les principales fonction de liaisons sont :

- la fonction logit :  $G(p) = \text{logit}(p) = \ln(1/(1-p))$  que l'on utilise dans le modèle dit logistique ;
- la fonction log-log :  $G(p) = \ln(-\ln(1 - p))$  ;
- l'inverse de toute fonction de répartition d'une loi continue sur  $\mathbb{R}$ . En particulier, l'inverse de la fonction de répartition de la loi normale, que l'on note traditionnellement  $\Phi^{-1}$ , est utilisée pour le modèle dit probit.

Le modèle logistique présente le gros avantage de permettre l'utilisation des odds-ratio qui représentent un outil très utile d'interprétation.

On estime alors les coefficients (qu'on nomme  $\hat{\beta}$ ) par maximum de vraisemblance.

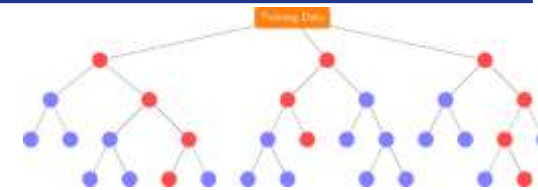
Pour des valeurs positives de  $\beta$ , plus  $\beta$  est grand, plus le régresseurs correspondant a un fort pouvoir de discrimination (une petite variation de ce régresseur peut occasionner une forte variation de  $p(x)$ ).

Mais cela doit être relativisé par le fait que la valeur de  $\beta$  (et donc de l'estimateur  $\hat{\beta}$ ) dépend fortement de l'unité de mesure utilisée ou de la modalité de référence choisie dans le cas de variables qualitatives.

**Dans notre cas, on prendra toujours la modalité la moins risquée en modalité de référence.**

# Les étapes de construction d'un score

## 6. Construction du modèle



### Démarche d'estimation du meilleur modèle logistique

Nous avons à notre disposition  $n$  variables candidates pour expliquer la donnée cible, toutes discrétisées dans les étapes précédentes.

Toutes ces données ne rentreront pas dans le modèle du fait des corrélations entre variables candidates qui viendront biaiser l'estimation du maximum de vraisemblance.

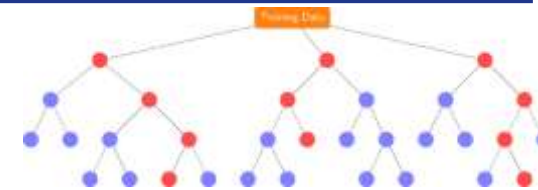
Ainsi, on doit **choisir quelles variables doivent entrer dans le modèle pour avoir le modèle le plus performant possible**.

Nous allons lister dans les slides suivantes diverses méthodologie pour sélectionner les « meilleures » variables et pour obtenir le modèle le plus performant.

**A partir d'ici, l'étape de construction du meilleur modèle est réalisée sur l'échantillon d'apprentissage, tandis que les étapes de performances (Gini, AUC, ...) sont à réaliser sur les 3 échantillons en parallèle pour vérifier que le modèle à des performances stables et semblables sur tous les échantillons, et qu'il ne souffre donc pas de biais de surapprentissage.**

# Les étapes de construction d'un score

## 6. Construction du modèle



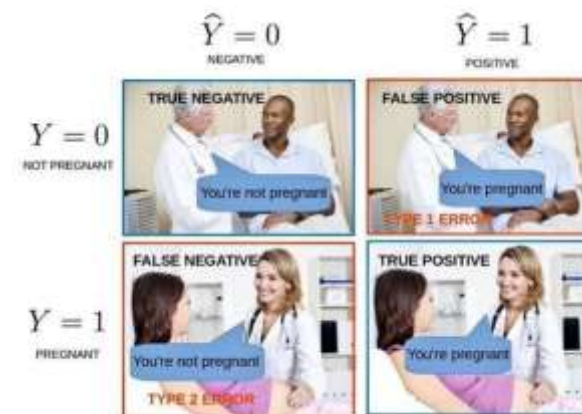
### Comment mesurer les performances d'un modèle ? Via ses prédictions

Lorsqu'un modèle de classification renvoi la probabilité d'appartenir à une classe, on peut alors convenir d'un seuil de probabilité sous lequel on considère que le modèle renvoi « 0 » et au dessus duquel il renvoi « 1 ». Par exemple, si le modèle renvoi une probabilité de « 43% », on classera la prédiction en « 0 » avec un seuil de 50% et en « 1 » avec un seuil de 40%.

A un seuil fixé, on peut alors construire une « matrice de confusion » en comparant la classe prédite par le modèle selon le seuil et la classe réellement observée.

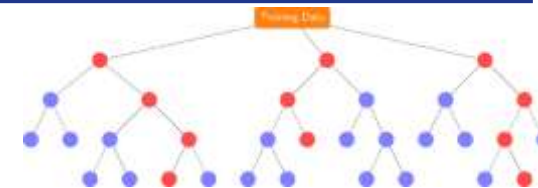
Toujours à seuil fixé, on peut alors calculer divers indicateurs importants, tels que :

- Précision : nombre de vrais positifs parmi tous les positifs prédits : exprime la part de bonne prédiction ;
- Recal (sensibilité) : nombre de vrais positifs sur le total de réellement positifs : exprime la part des positifs prédits par le modèle
- FPR (False Positive Rate)



# Les étapes de construction d'un score

## 6. Construction du modèle



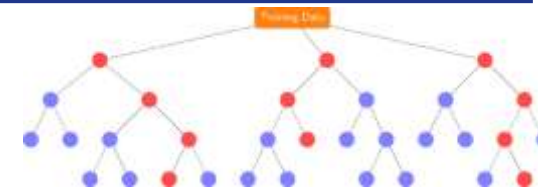
### Comment mesurer les performances d'un modèle ? Via ses prédictions

Voici une liste relativement exhaustive des métriques existantes à seuil fixé :  
[https://en.wikipedia.org/wiki/Sensitivity\\_and\\_specificity#Confusion\\_matrix](https://en.wikipedia.org/wiki/Sensitivity_and_specificity#Confusion_matrix)

		Predicted condition		Sources: [16][17][18][19][20][21][22][23] <a href="#">view</a> · <a href="#">talk</a> · <a href="#">edit</a>		
		Total population = P + N	Predicted positive (PP)	Predicted negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P) <sup>[a]</sup>	True positive (TP), hit <sup>[b]</sup>	False negative (FN), miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error <sup>[c]</sup> = $\frac{FN}{P} = 1 - TPR$	
	Negative (N) <sup>[d]</sup>	False positive (FP), false alarm, overestimation	True negative (TN), correct rejection <sup>[e]</sup>	False positive rate (FPR), probability of false alarm, fall-out type I error <sup>[f]</sup> = $\frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N} = 1 - FPR$	
Prevalence = $\frac{P}{P + N}$		Positive predictive value (PPV), precision = $\frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) = $\frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$	
Accuracy (ACC) = $\frac{TP + TN}{P + N}$		False discovery rate (FDR) = $\frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) = $\frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$	
Balanced accuracy (BA) = $\frac{TPR + TNR}{2}$		F <sub>1</sub> score = $\frac{2 \text{ PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2 \text{ TP}}{2 \text{ TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}{1}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$	

# Les étapes de construction d'un score

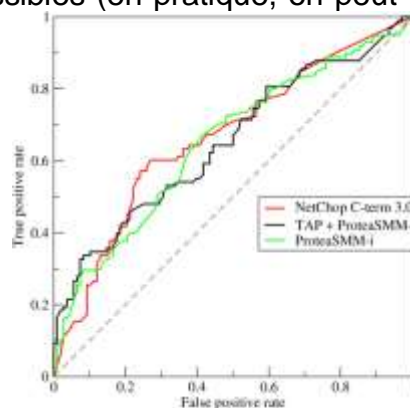
## 6. Construction du modèle



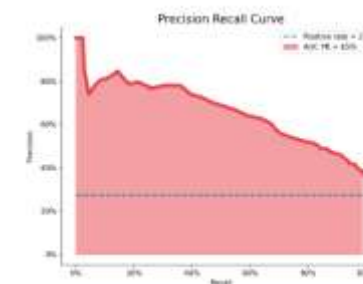
### Comment mesurer les performances d'un modèle ? Via ses prédictions

Si on calcule ces valeurs pour l'ensemble de seuils possibles (en pratique, on peut faire les 1000 seuils de 0% à 100% par pas de 0,1%), on peut alors construire des courbes d'intérêt :

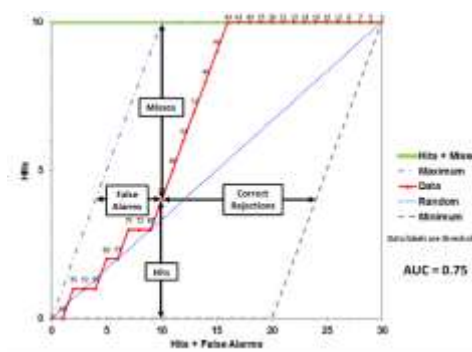
- **Courbe ROC** : courbe du TPR (Recall) en fonction du FPR ;



- **Courbe PR** : courbe du Precision en fonction du Recall ;

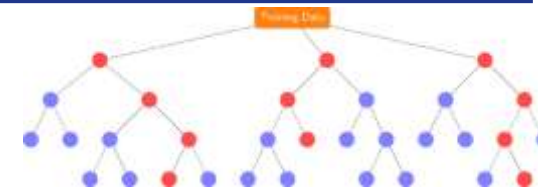


- **Courbe TOC** : courbe du TP en fonction du PP  
[https://en.wikipedia.org/wiki/Total\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Total_operating_characteristic)



# Les étapes de construction d'un score

## 6. Construction du modèle



### Comment mesurer les performances d'un modèle ? Via ses prédictions

On calcule ensuite l'**Aire Sous la Courbe (AUC)** pour chacune de ces courbes.

Note : on obtient toujours  $AUC-ROC = AUC = TOC$  par construction.

- L'**AUC-ROC** à l'immense avantage d'être une fonction de coût (loss function). Cela veut donc dire que l'on peut comparer 2 AUC entre elles, issues de 2 modèles et de deux jeux de données différents, et qu'on peut les classer.

Une AUC-ROC de 50% = modèle aléatoire.

Une AUC-ROC de 100% = modèle parfait.

Plus l'AUC augmente, meilleur est le modèle en termes de TPR sans dégrader les FPR.

On privilégiera donc les modèles avec une AUC forte.

**En pratique, on utilise plutôt le Gini =  $(2 \times AUC-ROC) - 1$ , et on cherche les modèles qui maximisent le Gini.**

En pratique, on cherchera des modèles ayant un Gini minimum de 60% (AUC-ROC de 80%), le seuil pouvant être abaissé dans certains cas très précis.

- L'**AUC-PR** n'est PAS une fonction de coût. On ne peut donc pas comparer des modèles issus de jeux de données différents entre eux et il n'existe pas de seuil de performance comme les 60% de Gini.

La précision du modèle aléatoire équivaut mathématiquement à la constante du taux de positifs ( $P/(P+N)$  = prévalence), et ce quelque soit le seuil.

On a donc **Taux de positifs  $\leq$  AUC PR d'un modèle de Machine Learning  $\leq$  100%**.

On mesurera alors aussi le **Ratio de Performance, qui vaut AUC-PR du modèle / AUC-PR du modèle aléatoire**.

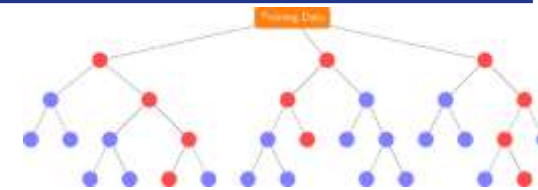
Cela permet de comparer entre eux des modèles issus de données distincts et avec des déséquilibres forts.

En effet, l'AUC-PR aura tendance à avoir de faibles valeurs si le jeu de données est déséquilibré. Le ratio de Performance est donc un bon indicateur complémentaire.



# Les étapes de construction d'un score

## 6. Construction du modèle



### Comment mesurer les performances d'un modèle ? Via ses prédictions

Il existe d'autres métriques basées uniquement sur les prédictions du modèle, comme le F1-score, pour lequel on peut calculer la courbe en fonction du seuil, ou, de façon générale, les F-beta-score.

#### La famille des F-beta scores

Le F1-score appartient à la famille plus large des **F-beta scores**. Dans le cas du F1-score, les erreurs (FN+FP) faites par le modèle sont pondérées par un facteur  $1/2$ . Le F1-score accorde ainsi la même importance à la précision et au rappel, et donc aux faux positifs et aux faux négatifs.

Le F-beta score permet de varier la pondération de ces termes :

$$F_{\beta}\text{-score} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Ce qui s'écrit également :

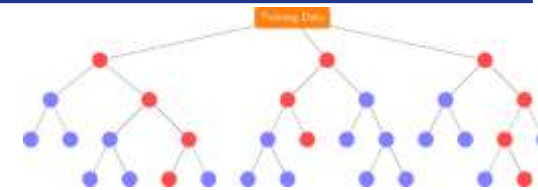
$$F_{\beta}\text{-score} = \frac{TP}{TP + \frac{1}{1+\beta^2}(\beta^2 FN + FP)}$$

Résumons :

- Pour  $\beta \geq 1$ , on accorde **plus d'importance au recall** (autrement dit aux faux négatifs).
- Pour  $\beta \leq 1$ , on accorde **plus d'importance à la précision** (autrement dit aux faux positifs).
- Pour  $\beta = 1$ , on retrouve le F1-score, qui accorde autant d'importance à la précision qu'au recall.

# Les étapes de construction d'un score

## 6. Construction du modèle



### Comment mesurer les performances d'un modèle ? Via sa vraisemblance

Il existe d'autres façon de mesurer la qualité d'un modèle sans prendre en compte la précision de ces prédictions. On peut utiliser la vraisemblance du modèle.

Pour cela, on utilise les critères :

- AIC : Le critère d'information d'Akaike s'écrit comme suit :

$$AIC = 2k - 2 \ln(L)$$

où  $k$  est le nombre de paramètres à estimer du modèle et  $L$  est le maximum de la [fonction de vraisemblance](#) du modèle.

- BIC : Il s'écrit :

$$BIC = -2 \ln(L) + k \cdot \ln(N)$$

avec  $L$  la [vraisemblance](#) du modèle estimée,  $N$  le nombre d'observations dans l'échantillon et  $k$  le nombre de paramètres libres du modèle<sup>1</sup>.

On privilégiera toujours un modèle avec un AIC/BIC faible à un modèle avec un AIC/BIC élevé par principe de parcimonie.

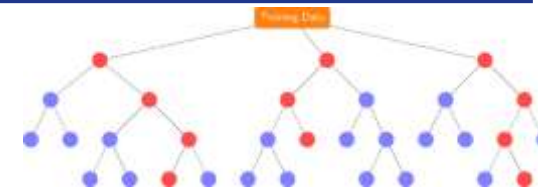
En effet, l'ajout de paramètres à un modèle augmente sa vraisemblance et sa qualité prédictive.

En particulier donc, **l'ajout de nombreuses variables dans un modèle augmentera de fait sa qualité, au détriment de sa parcimonie.**

**La question est donc : comment choisir le nombre de variables à faire rentrer dans mon modèle et quelles sont ces variables pour que le modèle soit précis (via les mesurer de la qualité des prédictions) et parcimonieux (via les critères AIC/BIC et avec un nombre de variables raisonnable pour justifier de sa bonne interprétabilité?)**

# Les étapes de construction d'un score

## 6. Construction du modèle



### Démarche d'estimation du meilleur modèle logistique

Nous avons à notre dispositions n variables candidates pour expliquer la donnée cible, toutes discrétisées dans les étapes précédentes.

Toutes ces données ne rentreront pas dans le modèle du fait des corrélations entre variables candidates qui viendront biaiser l'estimation du maximum de vraisemblance.

Ainsi, on doit **choisir quelles variables doivent entrer dans le modèle pour avoir le modèle le plus performant possible, tout en restant parcimonieux** (de façon générale, on essaye de limiter à une dizaine de variable maximum).

1. On commence par construire de façon automatique des modèles par méthodes de sélection forward, backward et stepwise ;
2. Ensuite, on construit n-1 modèles via un processus itératif en partant du modèle à 1 variable et en ajoutant les variables une à une dans l'autre de liaison avec la donnée cible).

On doit vérifier pour chaque variable (et donc notamment à chaque fois qu'une variable est ajoutée dans les modèles itératifs, mais aussi au global dans les modèles automatiques) que :

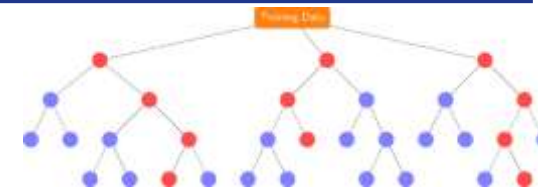
- **Les coefficients de chaque modalité sont significatifs** (5% en général, 10% accepté si nécessaire) ;
- La **significativité globale** des coefficients est ok (5%) ;
- **Les coefficients sont croissants avec le risque de la modalité** (par construction, la modalité 2 et moins risquée que la 3. Le coef de la modalité 3 doit donc être plus grand que celui de la 2) ;
- Les **odds ratios sont significatifs** entre chaque modalités d'une même variables (positifs et que leur IC ne contiennent pas 1).

Si un de ces critères n'est pas vérifié, on ne conserve pas le modèle et on supprime la variable mise en cause pour la suite (et donc on ne l'essaye pas dans les modèles suivants).

En particulier, si une modalité uniquement n'est pas significative, on peut tester un nouveau découpage.

# Les étapes de construction d'un score

## 6. Construction du modèle



### Démarche d'estimation du meilleur modèle logistique

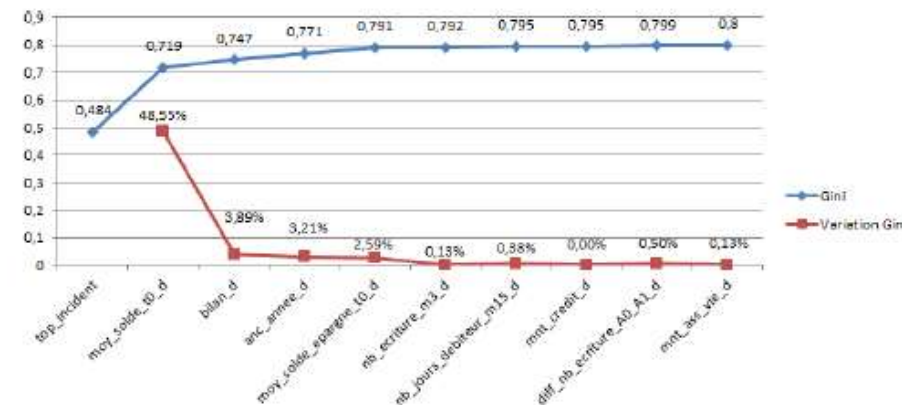
Pour chaque modèles (3 automatiques et n-1 itératifs), on compare les **performances de qualité du modèle** (AIC, BIC, AUC, Gini, autre métrique, ...).

On conserve le modèle minimisant l'AIC et le BIC et maximisant le Gini.

Si un tel modèle n'existe pas, on conserve le modèle minimisant l'AIC, le modèle minimisant le BIC et celui maximisant l'AUC (et donc le Gini).

On pourra alors les comparer par d'autres critères (qui ne sont pas des fonctions de coûts, donc à ne pas optimiser dans l'algorithme contrairement au Gini ou à l'AIC/BIC) comme :

- AUPRC : comparer le *ratio de performance*, qui est le taux de variable cible dans la base x AUCPR ;
- L'aire sous la courbe F1-score – treshold (ou autre F\_beta score) ;
- Parcimonie du modèle : à performances égales ou presque, on privilégiera le modèle avec le moins de variable explicatives ;
- R<sup>2</sup> du modèle ;
- Etc...



Vous pouvez proposer les métriques de votre choix, tant que le Gini reste un indicateur de premier plan de votre décision du meilleur modèle et que chaque nouvel indicateur est bien présenté.

**Pour le projet, il faudra présenter la comparaison d'au moins 2 modèles logistiques et présenter les métriques qui ont permis la décision entre les deux modèles.**

# Les étapes de construction d'un score

## 7. Restitution du modèle

### Restitution du meilleur modèle

Une fois le « meilleur » modèle sélectionné aux étapes précédentes, après analyse de ces performances (Gini, AUC-PR, AIC, BIC) et de sa parcimonie, il faut être en mesure de restituer le modèle afin de justifier de son explicabilité.

Pour cela, et c'est là aussi tout l'avantage des modèles linéaires, on restitue le modèle sous la forme de « grille de score ».

Pour passer des valeurs des coefficients  $\beta$  de chaque modalité à un score sur 1000, on calcule les pondérations associées aux modalités des variables dans le but de créer une échelle de score (allant de 0 à 1000 en pratique). 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.

On calcule également des niveau de contribution de chaque variable (contribution à l'échelle et contribution au score).

Nom de la variable	Nom technique des variables	Modalités	Population	Tx de défaut	Ecart relatif	Coefficient $\beta$	Points /1000	Contribution de la variable
Age et activité de la PP	date_naissance_pp date_observation top_emploi	1	<=17ans et 364 jours	12,20%	0,03%	75,00%	0	19,70%
		2	18ans et 0 jours <= x <= 20ans et 364 jours	4,66%	0,12%	87,76%	0,2297	
		3	PP retraitée OU >= 62ans et 0 jours	32,88%	0,98%	57,58%	0,3672	
		4	21ans et 0 jours <= x <= 61ans et 364 jours & PP travaille	50,26%	2,31%		0,8467	
Détenction de crédits immobiliers	nb_credits_immo	1	Au moins 1 crédit immobilier	20,10%	0,09%	89,66%	0	14,10%
		2	Pas de crédit immobilier	79,90%	0,87%		0,673	
Présence d'arriérés significatifs sur les 6 derniers mois	nbj_arr_sig anciennete_relation	1	Total de 0 jours d'arriérés sur 6 mois	70,01%	0,07%	41,67%	0	13,40%
		2	1 <= total de jours d'arriérés sur 6 mois <= 9	20,08%	0,12%	20,00%	0,3156	
		3	Pas assez de recul temporel	3,41%	0,15%	94,81%	0,4136	
		4	10 <= total de jours d'arriérés sur 6 mois <= 30	3,17%	2,89%	24,35%	0,8126	
		5	Au moins 31 jours d'arriérés sur 6 mois	3,33%	3,82%		1,23	



Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## 7. Restitution du modèle

### Calcul des pondérations /1000

Le calcul des pondérations se fait de la façon suivante est décrit en annexe :

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  $\alpha_j = \max[c(j, i)]$  le coefficient maximum pour la variable  $j$ .

On pose finalement :

$$SC(j, i) = 1000 \times \frac{|c(j, i) - \alpha_j|}{\sum_j \max_i c(j, i)}$$

Cette formule n'est valable que dans la situation où tous les coefficients estimés sont positifs et donc  $\min_i c(j, i) = 0, \forall j$ . D'où l'intérêt d'avoir mis les modalités les moins risquées en référence lors de la régression logistique.





Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## 7. Restitution du modèle

### Calcul des contributions échelle

La contribution d'échelle  $CTR_j$  de la variable  $j$  se calcule très simplement :

$$CTR_j = \frac{\max_i SC(j, i)}{10}$$

### Calcul des contributions score

La contribution  $q_j$  au score de la variable  $j$  se calcule à partir des pondérations (notes) calculées précédemment. On a :

$$q_j = \frac{\sqrt{\sum_{k=1}^m p_k (SC(j, k) - \overline{SC_j})^2}}{\sum_{i=1}^n \sqrt{\sum_{k=1}^m p_k (SC(j, k) - \overline{SC_i})^2}}$$

où

$p_k$	: part de la population échantillonnée 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.

# Les étapes de construction d'un score

## 7. Restitution du modèle

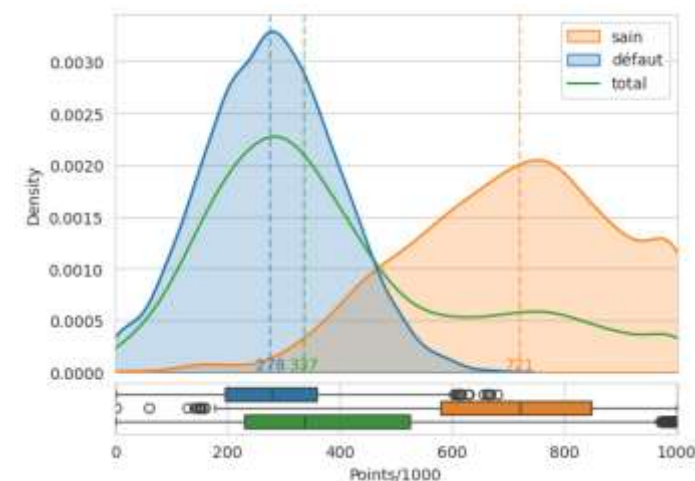
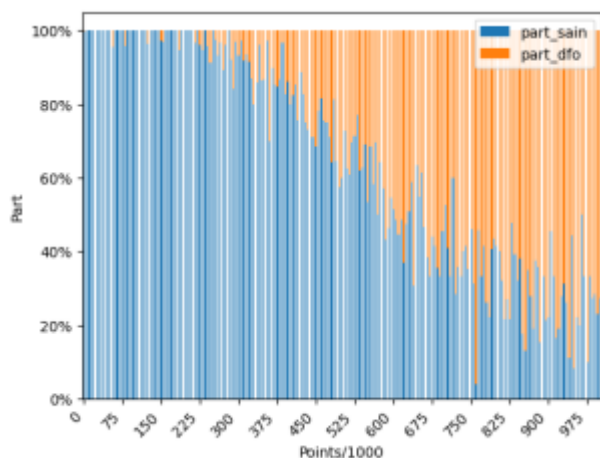
### Restitution du meilleur modèle

Les étapes précédentes permettent d'une part la restitution simple et lisible du modèle linéaire logistique sous forme de grille facilement interprétable par n'importe quel métier de la banque et par le régulateur (BCE, EBA, ACPR, Banque de France), et d'autres parts de convertir la note de régression, qui pour rappel retourne une probabilité, à une note /1000, avec 0 étant le pire score pour les clients les plus risqués et 1000 étant le meilleur score pour les clients les moins risqués d'après le modèle.

Cette nouvelle note / 1000 sera plus pratique pour la suite des travaux.

En effet, on peut d'une part facilement analyser les courbes de densité des clients cible ou non selon les points que leur attribue le modèle. C'est un outil visuel très pratique pour se rendre compte de la qualité du modèle (en plus des techniques mentionnées plus tôt, AUC, ...).

On en profite aussi pour regarder, pour chaque note de score, la part de client sain ayant cette note et la part de client cible.



# Les étapes de construction d'un score

## 7. Restitution du modèle

### Restitution du meilleur modèle

En théorie, le score est construit et on pourrait s'arrêter ici, on considérant qu'on a bien une note/1000 pour chaque client, qui définit le risque.

En pratique, on construit une échelle de risque associée à ce score.

Il s'agit de regrouper ensemble des clients ayant des contextes de risque proches d'après le score, et de leur attribuer une note de risque commune.

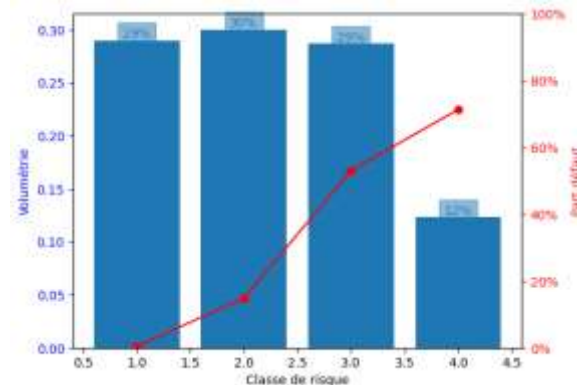
Pour cela, on construit des Classes Homogènes de Risque, ou CHR.

Il s'agit en fait de discrétiser la note de risque /1000 en n modalités, avec les mêmes contraintes que pour les variables à discrétiser plus tôt, c'est-à-dire :

- le taux du critère à modéliser entre les modalités d'une variable doit être croissant (la modalité 1 à le plus faible taux, la nième à le plus fort taux) ;
- Les modalités doivent être stables en taux du critère à modéliser (pas de croisement entre les modalités au cours du temps) et en volumétrie ;
- Un **effectif minimum de 5% 1%** de la population doit être trouvé dans chaque modalité (à l'exception, de variable particulièrement explicative du critère à modéliser et particulièrement fiable, pour exemple, dans le cadre de score de risque, pour des variables de dépassement d'autorisation ou d'impayé un effectif moindre est toléré) ;
- Les modalités doivent être séparée d'au moins **30% relatif** en termes de taux de variable cible.

A l'aide des graphiques précédents, on cherche des bornes qui permettent de regrouper les individus en CHR.

L'objectif est en général d'avoir 10 CHR, qui discriminent le défaut autant que possible.





Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## 7. Restitution du modèle

### Restitution du meilleur modèle

Enfin, il convient de valider la qualité du découpage en CHR.

La encore, on vérifie les stabilité des classes en volume et en risque au cours du temps.

Diverses autres analyses peuvent être réalisées, comme :

- Les matrices de migrations temporelles entre CHR ;
- La performance temporelle du score ;
- La courbe LIFT ;
- ...



Ma vie. Ma ville. Ma banque.

# Les étapes de construction d'un score

## Comment construire un score?

### Comment construire un score

Les étapes principales de la construction d'un score sont :

1. Définition des objectifs et périmètre d'étude
2. Inventaire et collecte des données
3. Construction de la base d'analyse (dont data splitting et création de variables)
4. Sélection des variables
5. Construction de modalités
6. Construction du meilleur modèle (focus important attendu sur les métriques de décision du meilleur modèle)
7. Construction du score /1000 à partir du modèle
8. Construction de CHR



Ma vie. Ma ville. Ma banque.

## 1 Présentation de LCL - Modélisation

## 2 Méthodologie de construction d'un score

## 3 Présentation du projet

## 4 Scores concurrents







Ma vie. Ma ville. Ma banque.

# Le projet

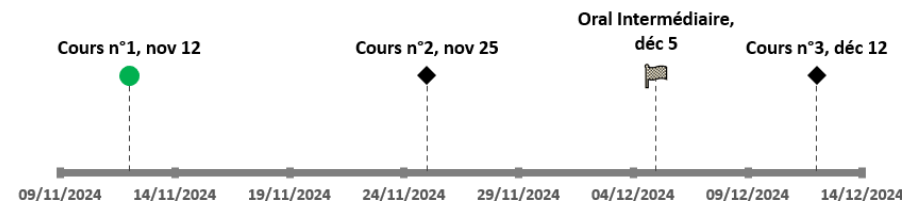
## Attendus et livrables

### Objectif

L'objectif du projet scoring 2024-25 est de réaliser un **score** permettant d'obtenir la **probabilité de défaut** sur la population des **clients professionnels** de LCL, sur un sous segment de votre choix..

Il conviendra de livrer :

- **1 modèle de score par méthodologie classique** (voir les slides suivantes) ;
- Un **document Word expliquant étape par étape la méthodologie suivie pour la construction du score** :
  - Le document peut/doit s'inspirer (voir suivre) de la méthodologie présentée dans ces slides ;
  - Le document doit être précis, chaque choix doit être détaillé, et en particulier les choix pouvant paraître anodins (variable non utilisée, segmentation des variables retenues, ...) ;
  - Une **attention particulière sera portée aux méthodes d'estimation des performances du modèle** (comparaisons, explications, limites et avantages de chacune)
- **1 modèle concurrent par méthodologie alternative** :
  - Sur ce modèle concurrent, on attendra peu de description quant à la méthodologie de construction du modèle ;
  - Des analyses de **comparaisons des performances entre les deux modèles** sont attendues ;
  - Des tentatives d'explicabilité du modèle alternatif, et la comparaison de l'explicabilité des deux modèles serait un plus.



La **présentation orale intermédiaire** du 05/12 portera uniquement sur le score classique. Il conviendra de présenter synthétiquement les étapes suivantes :

- Sélection des variables et construction des modalités (1/2 exemples max) ;
- Construction du modèle ;
- Performances du modèle ;
- Grille de score.



Ma vie. Ma ville. Ma banque.

# Le projet

## Note et bonus

### Objectif

Le projet sera noté de la façon suivante :

- **Note sur la présentation orale intermédiaire :**

- si, au moment de la présentation, aucun résultat n'est disponible, un malus sur la note finale sera appliqué ;
- en revanche, si vous avez essayé de faire quelque chose de propre mais que vous avez pris une mauvaise voie, la présentation servira pour vous réorienter, et aucun malus ne sera appliqué (une bonne note est tout à fait possible avec un modèle erroné mais bien présenté, qui suit les démarches présentées dans la suite du cours pour cet oral) ;

- **Note sur le rapport final :**

- La qualité de la méthodologie de construction du modèle est plus importante que les performances du modèle en lui-même!
- La qualité des codes ne sera pas évaluée, mais des codes restitués « *salement* » seront sanctionnés.
- La présence d'un modèle concurrent est requise, ainsi qu'une comparaison des performances.

- **Bonus:**

- Une base de type Kaggle (base sans la donnée cible) vous sera également fournie. Sur chaque segment modélisé, l'équipe obtenant le meilleur Gini sur la base Kaggle obtiendra un bonus de point.
- Des classements intermédiaires seront réalisés :
  - À la fin de la 2<sup>ème</sup> séance de cours ;
  - Après l'oral ;
  - A la fin de la 3<sup>ème</sup> séance.
- Seule la soumission finale (au moment de la livraison du rapport) fera foi. Les autres soumissions sont indicatives.
- Pour soumettre, envoyer un fichier csv par mail au format suivant :
  - Id / probabilité de 1 (colonne renommée p1)
  - Ne mettre aucune autre donnée.

id	p1
1	0,025
2	0,946
3	0,521
...	...



Ma vie. Ma ville. Ma banque.

# Le projet

## Présentation du périmètre

### Périmètre

Chez LCL, **les clients sont notés tous les mois quand à leur propension à faire défaut à horizon 12 mois.**

La notion de défaut est une notion réglementaire extrêmement précise et encadrée, définit par la BCE et l'EBA dans le règlement européen encadrant les réglementation Bâlois de gestion des risques bancaires (plus précisément, on parle ici de l'article 178 du Règlement (UE) n°575/2013, dit CRR. Pour les plus férus, la réglementation est détaillée dans les Guidelines EBA/GL/2016/07).

Pour faire simple, un client est déclassé en défaut bâlois lorsqu'il ne respecte plus ses engagements contractuels de crédits et commencent donc à ne plus rembourser une partie ou la totalité de ses mensualités.

En pratique, un client doit être déclassé en défaut lorsqu'il remplit au moins l'une de ces conditions :

- Le client a des **arriérés significatifs** depuis plus de 90 jours consécutifs (la notion de significativité et la granularité du calcul des arriérés au niveau débiteur sont expliquées dans la réglementation et ne font pas l'objet de ce cours) ;
- Le client présent un critère dit **d'UTP** (« Unlikelihood To Pay » - probablement absence de paiement). Parmi ces critère d'UTP, on retrouve :
  - le fait d'être considéré comme « surendetté » auprès de la banque de France (pour un client particulier) ;
  - Le fait d'être en procédure collective (redressement judiciaire, liquidation judiciaire, etc... pour un client professionnel) ;
  - Le fait d'être géré dans un service de recouvrement contentieux ;
  - Le fait d'avoir du modifier en urgence ces obligations de crédits afin de pouvoir continuer à les rembourser ;
  - Et autres non détaillées ici (comme la contagion notamment, ou les perspectives négatives).

Cette note de risque prend la forme d'une probabilité de défaut calculée pour chaque client.

Celle-ci est utilisée au quotidien dans la banque, à l'octroi comme au cours de la relation avec le client, pour obtenir un prêt, renégocier son crédit, renégocier ses tarifs commerciaux, etc...



Ma vie. Ma ville. Ma banque.

# Le projet

## Présentation du périmètre

### Périmètre

**Pour avoir une vision à date du risque d'un client à venir sur les 12 prochains mois, on utilise un score de risque!**

Chez LCL, on distingue le portefeuille en plusieurs catégories de clients, selon le « marché » du client :

- **Le marché Retail** : c'est la banque de détail. On retrouve plusieurs sous-marchés :
  - Le **marché des particuliers** : c'est les clients Personnes Physiques comme vous et moi, qui ont une relation bancaire pour leur gestion quotidienne (compte courant, CB, crédit immobilier, crédit à la conso, etc...) ;
  - Le **marché des professionnels** : c'est les clients Personnes Morales (c'est-à-dire des TPE/PME, sociétés ou associations) de petite taille jusqu'à taille intermédiaire, ainsi que des clients PP Entrepreneurs Individuels. La relation avec la banque est double : gestion au quotidien des liquidités et comptes de la banque, et financement des projets, matériels, ...
- **Le marché Corporate** : c'est la banque des (grosses) Entreprises et des multinationales. Ces clients sont multi bancarisés et cherchent des solutions de financement précises pour des besoins de plus grande envergure.

Peu importe le marché considéré, on utilise à chaque fois des scores de risque pour modéliser le risque de chaque client.

En particulier, sur le marché des professionnels, on retrouve une multitude de sous-segmentations possibles du portefeuille par secteur d'activités.

En effet, les clients « Professions Libérales » ne portent pas le même risque que les clients « SCI » ou « Artisans », ou encore « Pharmacie ». Chaque secteur d'activité évolue au sein d'une conjoncture macroéconomique qui lui est propre.



Ma vie. Ma ville. Ma banque.

# Le projet

## Présentation du périmètre

### Périmètre

**Chez LCL aujourd'hui, un unique score de risque est utilisé pour noter le risque de TOUS les clients professionnels.**

Ce score est très performant sur le portefeuille pris dans son ensemble.

Cependant, les performances varient assez largement d'un sous-périmètre à un autre, car les règles de décisions retenues, uniques pour tous les périmètres, ne sont pas toujours aussi bien adaptées à chaque périmètre.

L'objectif du score est donc de construire un score de risque, dont la variable cible est le défaut à horizon 12 mois, uniquement sur certains sous périmètres précis et distincts.

Nous vous proposons de travailler, aux choix, sur :

- **Les professions libérales de santé et pharmacie ;**
- **Les professions libérales hors santé ;**
- **Les SCI ;**
- **Les entreprises de commerce.**

Des bases de données vous seront transmises, avec toutes les données nécessaires à la réalisation du projet, ainsi qu'un dictionnaire de données.

Vous trouverez, dans le fichier excel :

- Feuille 1 : les données à utiliser pour l'apprentissage / le test du modèle (données clients entre Janvier et Juin 2023, avec défaut observé sur 12 mois jusqu'à Juin 2024) ;
- Feuille 2 : les données à utiliser pour l'analyse hors temps (données clients entre Juillet et Aout 2023, avec défaut observé sur 12 mois jusqu'à Aout 2024) ;
- Feuille 3 : les données de type Kagle à utiliser (mélange de données clients des mois de Janvier de 2011 à 2021, sans les données de défaut).



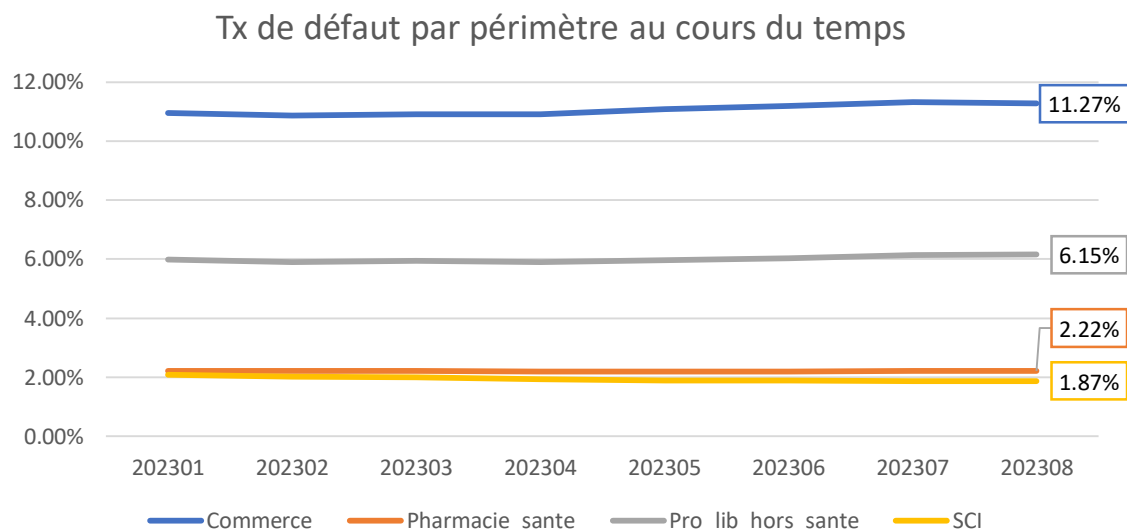
Ma vie. Ma ville. Ma banque.

# Le projet

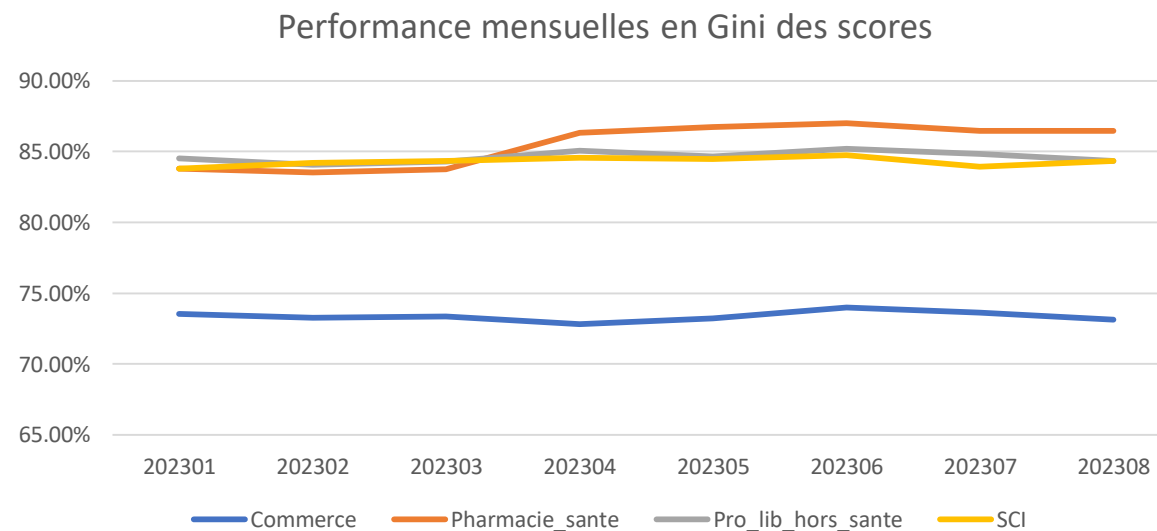
## Présentation du périmètre

### Taux de défaut mensuel par périmètre

Voici les taux de défauts mensuels par périmètre :



Voici les performances en Gini des scores au cours du temps







Ma vie. Ma ville. Ma banque.

## 1 Présentation de LCL - Modélisation

## 2 Méthodologie de construction d'un score

## 3 Présentation du projet

## 4 Scores concurrents





Ma vie. Ma ville. Ma banque.

---

# Scores concurrents

## Typologies de modèles utilisables

---

### Les modèles de classification

---

Pour le modèle concurrent, il conviendra toujours de construire un modèle dont l'outcome est une probabilité d'appartenir à la classe 1.

**Éléments**

**Complémentaires**

# Code

## 3. Construction de la base d'analyse



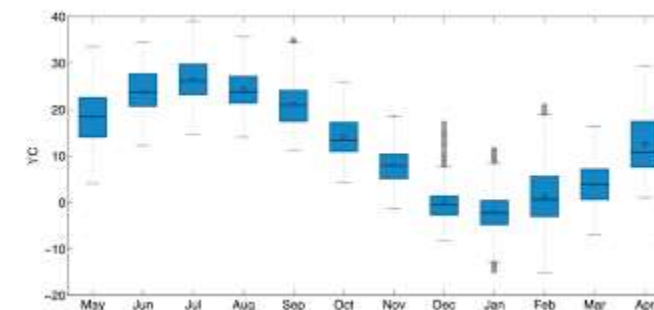
### ➤ Code pour obtenir le boxplot mois par mois pour vérifier la stabilité d'une variable

A partir d'un df de la forme :

On fait le code :

```
# On fait un boxplot mois par mois sur le sample total
fig, ax = plt.subplots()
sns.boxplot(data=sample, x='mois', y='score', ax=ax)
ax.set(xlabel='Mois', ylabel='Points/1000')
plt.show()
```

	Variable	mois
0	617.525362	10
1	277.163877	7
2	95.672395	9
3	420.422980	4
4	427.625981	10



# Code

## 7. Restitution du modèle

Exemple de code pour obtenir les courbes de densité séparées :

1. On crée des jeux de données labélisées pour tous les clients avec leur label (sain ou défaut) et le score/1000 calculé ;
2. On crée deux autres jeux nommés sample\_total et sample\_full

```
[25] # On copie sample et change tous les labels en 'total'
sample_total = sample.copy()
sample_total['label'] = 'total'

# On fusionne sample et sample_total
sample_full = pd.concat([sample, sample_total])
```

3. On plot

```
fig, ax = plt.subplots(2)
# ...
# On plot une kde de sample selon le label, avec le kde qui s'arrête à 0 et 1000
sns.kdeplot(data=sample, x='score', hue='label', common_norm=False,
            fill="true", ax=ax['A'])

# On plot en arrière plan un kde plot de sample_full
sns.kdeplot(data=sample_full, x='score', hue='label', common_norm=False,
            ax=ax['A'], legend=False)

# On rajoute le 'total' à la légende
ax['A'].legend(labels=['sain', 'défaut', 'total'])

ax['A'].set(xlim=(0, 1000))
# On supprime les tick des abscisses
ax['A'].set_xticks([])
```


```
# On supprime les tick des abscisses
ax['A'].set_xticks([])

# On plot un boxplot de sample selon le label
sns.boxplot(data=sample_full, x='score', hue='label', legend=False, ax=ax['B'])
ax['B'].set(xlim=(-2, 1001))

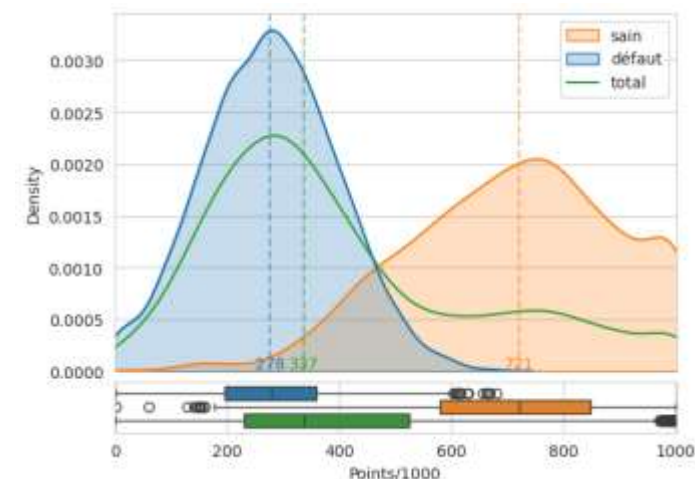
# On affiche les valeurs médianes par label sans qu'ils ne dépassent le pic du kde
for i, label in enumerate(sample['label'].unique()):
    median = sample[sample['label'] == label]['score'].median()
    ax['A'].axvline(x=median, color=f'C{i}', linestyle='--', alpha=0.5)
    # On écrit la valeur
    ax['A'].text(median, 0, f'(median:0f)', ha='center', va='bottom', color=f'C{i}')

# Idem pour le total
median = sample_full['score'].median()
ax['A'].axvline(x=median, color='C2', linestyle='--', alpha=0.5)
ax['A'].text(median, 0, f'(median:0f)', ha='center', va='bottom', color='C2')

ax['B'].set(xlabel='Points/1000')
sns.set_style('whitegrid')
plt.show()
```

0 s  sample.head()

	score	label
0	346.817420	sain
1	416.307354	sain
2	579.740995	défaut
3	795.688429	défaut
4	483.407771	sain



# Code

## 7. Restitution du modèle

Exemple de code pour obtenir la part de sain/défaut par note:

1. On crée un jeu de donnée pour tous les clients avec leur label (sain ou défaut) et le score/1000 calculé ;
2. On arrondi le score à la cinquième la plus proche et on calcule les parts de sains/défaut

```
# Pour chaque note /1000, on calcule la part ou pourcentage de sain et la part de défaut

# On arrondi le score au 5 le plus proche
sample['score_round'] = np.round(sample['score'] / 5) * 5

# Pour chaque note de score round, on regarde la part de sain et la part de défaut
# df vide initialisé
part_df = pd.DataFrame(columns=['score_round', 'part_sain', 'part_dfo'])

# Boucle sur les scores
for score in sample['score_round'].unique():
    sub_df = sample[sample['score_round'] == score]
    # On calcule la part de sain
    part_sain = sub_df[sub_df['label'] == 'sain'].shape[0] / sub_df.shape[0]
    # On calcule la part de défaut
    part_dfo = sub_df[sub_df['label'] == 'défaut'].shape[0] / sub_df.shape[0]

    # On met dans le part_df
    part_df = pd.concat([part_df, pd.DataFrame({'score_round': [score], 'part_sain': [part_sain], 'part_dfo': [part_dfo]})])

# On ordonne part_df
part_df = part_df.sort_values(by='score_round')
```

3. On plot

```
# On plot le graphique de part_df en bar empilés
fig, ax = plt.subplots()

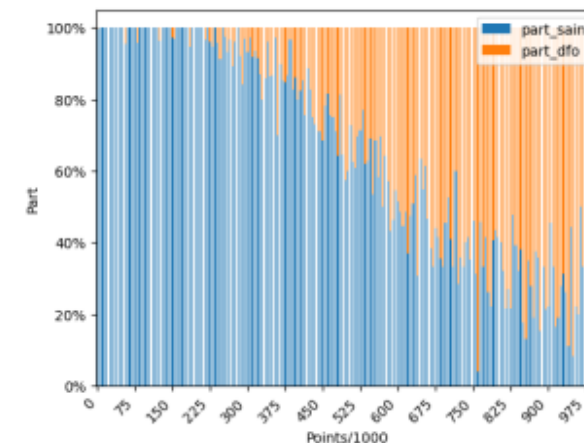
part_df[['part_sain', 'part_dfo']].plot(ax=ax, kind='bar', stacked=True,)

# On écrit les yticks en pourcentage
ax.set_yticklabels(['{:.0f}%'.format(x*100) for x in ax.get_yticks()])
# On écrit les xticks en diagonale
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, ha='right')
# On réécrit les xticks car ils sont illisibles
ax.set_xticklabels([str(int(x)) for x in part_df['score_round']])
# On affiche uniquement 1 xtick sur 5
ax.xaxis.set_major_locator(plt.MaxNLocator(20))

ax.set(xlabel='Points/1000', ylabel='Part')
plt.show()
```

sample.head()

	score	label
0	644.531297	sain
1	854.842965	défaut
2	789.603586	sain
3	409.758136	défaut
4	734.962540	défaut





# Code

## 7. Restitution du modèle

```
# On calcule la classe de risque
sample['classe'] = np.where(sample['score'] <= 200, 1,
                             np.where(sample['score'] <= 500, 2,
                                     np.where(sample['score'] <= 700, 3,
                                             np.where(sample['score'] <= 800, 3,
                                                     np.where(sample['score'] <= 900, 4,
                                                             np.where(sample['score'] <= 950, 4, 4))))))
sample
```

```
# On calcule la part de défaut par classe de risque
# df vide initialisé
chr_df = pd.DataFrame(columns=['classe', 'part_sain', 'part_dfo', 'vol_total'])

# Boucle sur les scores
for classe in sample['classe'].unique():
    sub_df = sample[sample['classe'] == classe]
    # On calcule la part de sain
    part_sain = sub_df[sub_df['label'] == 'sain'].shape[0] / sub_df.shape[0]
    # On calcule la part de défaut
    part_dfo = sub_df[sub_df['label'] == 'defaut'].shape[0] / sub_df.shape[0]

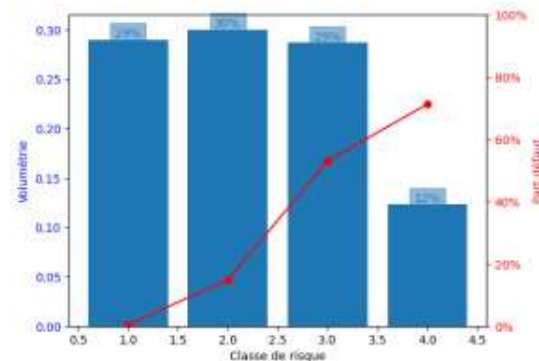
    # On calcule la volumétrie du total de sample qui est dans cette classe
    vol_total = sub_df.shape[0] / sample.shape[0]

    # On met dans le part_df
    chr_df = pd.concat([chr_df, pd.DataFrame({'classe': [classe], 'part_sain': [part_sain], 'part_dfo': [part_dfo], 'vol_total': [vol_total]})])

# On ordonne part_df
chr_df = chr_df.sort_values(by='classe')
chr_df
```

	score	label	score_round	classe
0	0.000000	sain	0.0	1
1	204.041692	defaut	205.0	2
2	771.404824	defaut	770.0	3
3	715.564291	defaut	715.0	3
4	170.854397	sain	170.0	1
...	...	...	...	...
6995	677.574890	defaut	680.0	3
6996	222.056693	defaut	225.0	2
6997	461.895569	sain	460.0	2
6998	285.182338	sain	285.0	2
6999	276.188154	sain	275.0	2

	classe	part_sain	part_dfo	vol_total
0	1	0.996059	0.003941	0.290000
0	2	0.851499	0.148501	0.300143
0	3	0.469124	0.530876	0.286857
0	4	0.286876	0.713124	0.123



```
# on plot la volumétrie sur l'ax y1
fig, ax1 = plt.subplots()
ax1.bar(chr_df['classe'], chr_df['vol_total'])
ax1.set_ylabel('Volumétrie', color='b')
ax1.tick_params('y', colors='b')

# on écrit la valeur de la volumétrie au dessus des barres dans un rectangle de couleur
for i, vol in enumerate(chr_df['vol_total']):
    ax1.text(i+1, vol, f'{vol*100:.0f}%', ha='center', va='bottom', color='b',
             bbox=dict(facecolor='b', edgecolor='b', alpha=0.5))

# on plot part dfo en ligne sur y2, ticks %
ax2 = ax1.twinx()
ax2.plot(chr_df['classe'], chr_df['part_dfo'], color='r', marker='o')
ax2.set_ylabel('Part défaut', color='r')
ax2.tick_params('y', colors='r')
ax2.set_ylim(0, 1)
ax2.set_yticks([0, 0.2, 0.4, 0.6, 0.8, 1])
ax2.set_yticklabels(['0%', '20%', '40%', '60%', '80%', '100%'])
# xlab
ax1.set_xlabel('Classe de risque')

plt.show()
```