Support Machine Vector
&
Machine Learning Interprétable

Présenté par :

Jenny
RAHARIMANANA

Kourosh
KAZEMI

LCL LE CRÉDIT LYONNAIS

MASTER ESA
Université d'Orléans

L’Histoire du Crédit Lyonnais (LCL)

1863

Naissance du Crédit Lyonnais par François Barthélemy & Henri Germain

1878

Le Crédit Lyonnais est la première banque de France

1945

Nationalisation de LCL

1999

Le Crédit Lyonnais est privatisée suite à un accord entre l'Etat et l'Union Européenne

2003

Rapprochement entre LCL et Crédit Agricole, qui devient la banque de Proximité

2018

"Ma vie. Ma ville. Ma Banque", LCL a l'ambition de devenir la 1ère banque urbaine de tous les habitants de la ville

I - Développement d’un modèle de Scoring

PLAN


01
Construction de la base d'analyse.
02
Sélection des variables.
03
Estimations et grille de score.
04
Analyse des performances.
05
Création de classes de risque.

Mise en contexte

Au coeur de la santé économique, les **institutions bancaires** constituent un pilier de la robustesse de la nation. C'est pourquoi ont lieu **diverses discussions internationales** afin de promouvoir des **règles de contrôles communes** afin d'éviter des **"dumping" bancaires**.

Pour faire face au mieux aux situations de **défaut,** la banque doit définir deux types de pertes : celles **attendues** (qui seront couvertes par les **provisions**) et celles **inattendues** (qui devront être couvertes par **les fonds propres**). Cette estimation de pertes inattendues constitue un des défis majeurs de la banque.

La banque peut, sous certaines conditions, estimer elle-même ses **paramètres de risque** afin de déterminer son **niveau de capitaux propres requis**.

Les accords de Bâle IV

Objectif : Exiger plus de capital pour couvrir la même quantité de risque.

Critères d'entrée en défaut

- Seuil d'arriérés franchis plus de 90 jours consécutifs
- Faillite / procédure judiciaire
- Restructurations pour risque avec perte économique supérieur à 1%
- Perspectives negatives d'experts
- Clients en défaut chez Interflimo
- Provisions sur le compte

Critères de sortie de défaut

- 3 mois après l'absence d'élément déclencheur de défaut pour les clients sans contrat RR
- 12 mois à débuter immédiatement dès la RR pour les clients avec contrat RR

Piste de travail

- L’objectif de notre projet est dans un premier temps d’effectuer une **régression logistique simple**, afin de pouvoir prédire au mieux **le risque de défaut bâlois sous la nouvelle discussion Bâle IV en perspective**.

- L’une des optiques principales de ce comité est de **renforcer les coussins de fonds propres** dont les banques disposent afin de pouvoir au mieux **faire face aux risques en cas de défaut**.

- C’est pourquoi, tout au long de notre projet, nous avons choisi **de minimiser autant que faire se peut le risque de type I**, c’est-à-dire **le risque d’attribuer un crédit à un individu risqué**, au dépend du risque de type II, c’est-à-dire le **coût d’opportunité** de laisser partir un « bon » client.

I – Construction de la base d’analyse

Nos variations

La variable cible

Objectif : Modéliser le risque de défaut de chaque individu de notre base

Indicateur du nouveau défaut Bâlois à 12 mois

Agencement des activités

Présentation de la base de données

Notre base contient **3 types de clients** :

Personne Privée

Il s'agit de **personnes physiques** dont les engagements appartiennent **au marché primaire**

Entreprise Individuelle

Il s'agit **de personnes physiques** dont les engagements appartiennent au **marché secondaire**.

Autre personne privée

Il s'agit de **personnes physiques** ayant **au moins un engagement** envers la banque.

Traitement des valeurs manquantes

- Notre base ne contient pas beaucoup de valeurs manquantes. En effet, c'est la variable **variation de l'indicateur de risque** qui en est le plus assujettie, avec environ **4,50% de valeurs manquantes**.

- Pour ne pas perdre d'information, nous avons choisi de ne pas supprimer les individus qui comportent des valeurs manquantes, mais plutôt de les traiter avec une **méthode d'imputation simple,** l'idée étant de **remplacer chaque valeur manquante par une donnée simulée avec une analyse portant sur tous les enregistrements**.

- Nous avons retenu la méthode du **Hot-Deck** dont le principe est le suivant : on remplace la **valeur manquante par une valeur observée chez un individu de notre base de données ayant les mêmes caractéristiques.**

- Cette méthode nous semblait plus pertinente que les méthodes d'imputation par médiane ou par Moyenne **non conditionnelles**, qui selon nous, ne faisaient pas assez appel aux caractéristiques de la base de données étudiée.

- Après avoir remplacer les données manquantes par cette méthode, nous avons vérifié que les statistiques de la nouvelle base étaient similaires à celles de notre base initiale. Nous nous sommes notamment penchés sur l'évolution de la **moyenne, de la médiane, de la variance, ainsi que des valeurs minimales et maximales**.

- Le résultat est concluant : on observe pour ces statistiques une variation **de moins de 5%** par rapport à celles de la base initiale.

Traitement des valeurs aberrantes

- Il existe plusieurs critères pour définir qu'un individu constitue une **valeur aberrante**. Pour notre étude, nous nous sommes penchés sur l'idée qu'une valeur est aberrante si **sa valeur est plus d'une fois et demie la longueur de l'écart interquartile**.

- Ces valeurs aberrantes ne représentent qu'environ **2,8 %** de nos données, nous avons donc choisi de les supprimer pour la suite de notre analyse.

- Par ailleurs, dans un soucis de pertinence des variables, nous avons choisi d'introduire deux variables dynamiques au sein de notre base de données :

1) **La variation du nombre de jours d'arriérés** : Cette variable nous permettra de capter si la situation d'un client s'est améliorée au cours des 3 derniers mois, si elle est restée la même ou si, au contraire, sa situation s'est dégradée. Cette variation reflète d'ailleurs, si l'on suppose les deux seuils franchis, l'un des critères d'entrée en défaut sous Bâle IV.

1) **La variation du montant de capital dû** : Cette variable nous permettra de capter si le client a été capable de rembourser ses engagements au cours des 3 derniers mois ou non. Cela nous semblait plus pertinent car la base de nous indique pas la date de souscritpion au prêt et donc complique l'interprétation du montant de capital dû en niveau.

Echantillonnage

- Pour la suite de notre étude, nous avons divisé notre échantillon en un sous-échantillon **d'apprentissage** sur lequel seront menés nos estimations et qui représentera environ **70% de notre base initiale**, ainsi qu'un sous-échantillon **test** sur lequel seront menés les évaluations de notre modèle et qui représentera environ **30% de notre base**.

- La division de notre échantillon doit se faire de telle sorte que la **représentatibilité** de notre base de données d'origine soit conservée dans chacun des sous-échantillons créés. Cela s'est fait naturellement grâce à **la procédure SURVEYSELECT sous SAS**.

200 000 individus dans le sous-échantillon APPRENTISSAGE

(70% de la base)

100 000 individus dans le sous-échantillon TEST

(30% de la base)

II- Sélection des variables

- Pertinence
- Non redondance
- Interprétabilité

Pré-selection des variables

CAS DES VARIABLES QUANTITATIVES

Etape 1 : Etude de la distribution de nos variables

Pour étudier les distributions de nos variables quantitatives, nous utilisons 3 tests, à savoir les tests de **Kolmogorov-Smirnov, de Cramer-von Mises et de Anderson-Darling.**

a. Test de Kolmogorov-Smirnov

Ce test repose sur les propriétés des **fonctions de répartitions empiriques** des variables étudiées. La statistique de test mesure **l'écart maximal entre la fonction de répartition empirique et la fonction de répartition à tester.**

Dans notre cas, on a :

H0 : Les variables sont distribuées normalement
H1 : Les variables ne sont pas distribuées normalement

CAS DES VARIABLES QUANTITATIVES

Etape 1 : Etude de la distribution de nos variables

Pour étudier les distributions de nos variables quantitatives, nous utilisons 3 tests, à savoir les tests de **Kolmogorov-Smirnov, de Cramer-von Mises et de Anderson-Darling.**

b. Test de Cramer-von Mises

Ce test permet de calculer l'équivalence entre la distribution empirique observée et la distribution théorique de référence fixée.

Dans notre cas, on a :

H0 : Les variables sont distribuées normalement
H1 : Les variables ne sont pas distribuées normalement

CAS DES VARIABLES QUANTITATIVES

Etape 1 : Etude de la distribution de nos variables

Pour étudier les distributions de nos variables quantitatives, nous utilisons 3 tests, à savoir les tests de **Kolmogorov-Smirnov, de Cramer-von Mises et de Anderson-Darling.**

c. Test d'Anderson-Darling

Cette statistique détermine dans quelle mesure les données suivent une loi de distribution spécifique. Plus la loi de distribution spécifiée (ici normale) sera ajustée à nos données, moins on aura tendance à rejeter notre hypothèse nulle.

Dans notre cas, on a :

H0 : Les variables sont distribuées normalement
H1 : Les variables ne sont pas distribuées normalement

Résultats

Pour les trois tests, on va systématiquement **rejeter fortement l'hypothèse nulle de normalité** de nos 41 variables quantitatives.

Nos variables ne sont donc pas normalement distribuées.

Pré-selection des variables

CAS DES VARIABLES QUANTITATIVES

Etape 2 : Etude de la significativité de nos variables

Pour étudier la significativité de nos variables quantitatives, nous nous sommes penchés sur un test statistique non paramétrique, à savoir le **test de Kruskal-Wallis.**

Le test de Kruskal-Wallis permet de déterminer si **les médianes de deux groupes diffèrent ou non,** auquel cas nous pourrons conclure que nos variables n'influent pas notre facteur de défaut**.** De manière plus approfondie, cette statistique de test permet de déterminer si des échantillons proviennent d'une même distribution.

Dans notre cas, on aura :

H0 : Les variables n'ont pas d'impact significatif sur la variable cible
H1 : Les variables ont un impact significatif sur la variable cible

Résultats

Toutes nos variables quantitatives ont obtenu une **p-value inférieure à 1% pour ce test**. Nous avons, donc dans tous les cas, fortement **rejeté l'hypothèse d'égalité des médianes du facteur de défaut** au sein de notre échantillon.

Nos variables sont donc toutes significatives pour la suite de l'étude.

Pré-selection des variables

CAS DES VARIABLES QUANTITATIVES

Etape 3 : Discrétisation de nos variables

Dans une optique d'optimisation de la grille de score et de regroupement d'individus, nous avons discrétisé nos variables à l'aide de la **méthode des déciles.**

L'idée est dans un premier temps de diviser notre échantillon en 10 déciles pour chacune des variables à discrétiser. Notre nombre de classe ayant été limité à 5 (expertise métier), nous avons donc **regroupé en une même classe les déciles voisins qui avaient la probabilité de défaut la plus similaire**.

Afin d'améliorer la qualité de notre discrétisation, nous avons utilisé deux méthodes : **un programme optimal purement statistique** et un **regroupement manuel qui s'apparente à l'expertise métier**. Nous avons ensuite combiné ces deux résultats pour définir nos classes.

Pré-selection des variables

CAS DES VARIABLES QUANTITATIVES

Etape 4 : Etude de la multicolinéarité de nos variables

La dernière étape consiste à vérifier la **multicolinéarité** de nos variables présélectionnées, c'est-à-dire s'il existe une **corrélation forte entre ces variables**.

En effet, si plusieurs variables dans le modèle capturent les mêmes effets, les coefficients de régression estimés ainsi que leur variance peuvent être élevés en valeur absolue ce qui entraîneraient par exemple une conclusion erronée compte aux tests statistiques usuels.

Pour étudier la colinéarité de nos variables quantitatives, nous avons procédé en deux étapes :

- l'étude des **corrélations globales** au sein de nos variables quantitatives
- l'étude du **V de Cramer** sur nos variables discrétisées pour déterminer quelle variable allait rester ou non dans notre étude.

Pré-selection des variables

CAS DES VARIABLES QUANTITATIVES

Etape 4 : Etude de la multicolinéarité de nos variables

a. Etude des corrélations globales :

Nous avons dans un premier temps étudié les corrélations globales au sein de toutes nos variables quantitatives. Nous avons estimé de manière arbitraire qu'une **corrélation** était jugée forte lorsqu'elle **dépassait ou s'approchait de 0,5 en valeur absolue**.

A partir des résultats, nous avons pu créer **9 groupes de variables corrélées entre elles**.

b. Choix des variables :

Au sein de nos 9 groupes de variables corrélées (discrétisées), nous avons effectué un **test du Chi-Deux** et récupérer **la valeur absolue du V de Cramer** pour déterminer quelle variable était la **plus corrélée à la variable cible** au sein de chaque groupe. Ce sont ces neuf variables et celles non corrélées qui ont été retenues pour la suite de notre analyse.

Récapitulatif

1. Etude de la distribution des variables (Kolmogorov, Darling, Mises) :
Aucune variable n'est gaussienne

3. Discrétisation des variables:
Méthode des déciles

5. Etude de la multicolinéarité :
Choix des variables non corrélées (V de Cramer)

2. Etude de la significativité (Kruskal-Wallis) :
Toutes les variables sont significatives

4. Etude de la multicolinéarité :
Calcul du coefficient de corrélation entre les variables

Pré-selection des variables

CAS DES VARIABLES QUALITATIVES

Etape 1 : Etude de la significativité de nos variables

Pour étudier la significativité de nos variables qualitatives, nous avons effectué un **test du Khi-Deux** et avons ensuite récupéré les **valeurs absolues des V de Cramer** de nos variables**.**

Toutes nos variables qualitatives ont obtenu une **p-value inférieure à 1% pour ce test**.

Nos variables sont donc toutes significatives pour la suite de l'étude.

Pré-selection des variables

CAS DES VARIABLES QUALITATIVES

Etape 2 : Recodage de nos variables

Certaines de nos variables qualitatives contenaient un nombre trop élevé de modalités. Nous avons donc été obligé de recoder les variables concernées.

Ce recodage s'est fait, comme pour le cas des variables quantitatives, à l'aide de la **méthode des déciles**. Ainsi, nous avons divisé notre échantillon en déciles, en fonction de la variable à étudier. Nous avons ensuite regroupé, au sein d'une même classe, les **déciles voisins qui partageaient le même niveau de risque de défaut**.

Pré-selection des variables

CAS DES VARIABLES QUALITATIVES

Etape 3 : Etude de la multicolinéarité de nos variables

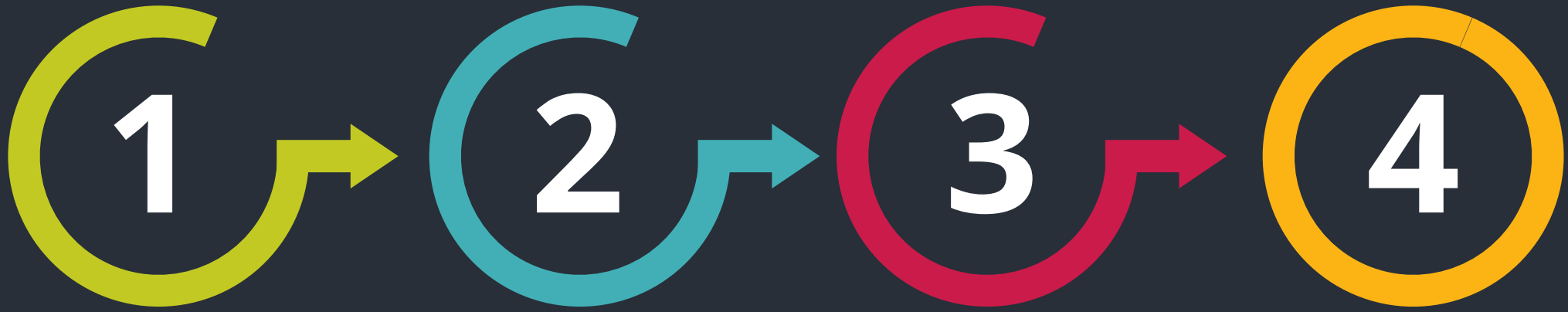
a. Etude des colinéarités globales :

Nous avons dans un premier temps étudié les relations globales au sein de toutes nos variables qualitatives. Nous avons estimé de manière arbitraire qu'une **colinéarité** était jugée forte lorsque **la valeur absolue du V de Cramer** associée **dépassait ou s'approchait de 0,5**.

b. Choix des variables :

Au sein de nos variables liées, nous avons effectué un **test du Chi-Deux** et récupérer **la valeur absolue du V de Cramer** pour déterminer quelle variable était la **plus corrélée à la variable cible** au sein de chaque groupe.

Récapitulatif



Etude de la significativité de nos variables (test du Chi-Deux) :
Toutes les variables sont significatives

Recodage de nos variables (méthode des déciles) :
Recodage de CODACVECO

Etude de multicolinéarité (V de Cramer)

Choix des variables non corrélées.

Pré-selection des variables

CAS DES VARIABLES QUALITATIVES ET QUANTITATIVES

Etude de la multicolinéarité de nos variables entre variables qualitatives et quantitatives

Nous cherchons ici à étudier la **colinéarité entre nos variables qualitatives et les variables quantitatives**. Pour se faire, nous allons directement utiliser nos variables recodées de manière à **comparer uniquement des variables qualitatives**.

Nous effectuons **un test de Chi-Deux** en croisant nos variables et nous récupérons ensuite **la valeur absolue de leur V de Cramer**. Nous avons fixé de manière arbitraire un **seuil de 0,5** au de-là duquel nous avons jugé une **colinéarité importante** entre les variables.

Pré-selection des variables

- Au terme de ces différentes procédures, nous avons décidé de pré-sélectionner **14 variables** dans notre modèle.
- Ces 14 variables répondent donc à tous les critères permettant de définir un bon modèle en terme de qualité d'ajustement. Notamment:

1. Nos variables sont **toutes fortement corrélées** à la variable cible.

1. Nos variables **ne sont pas corrélées** entre elles et ne **sont donc pas redondantes**.

1. Nos variables **sont significatives** et chacune de **leur modalité** le sont également. En cela, elles possèdent toute un **rapport de côte** dont l'intervalle de confiance **exclue la valeur unité**.

1. Nos variables sont **interprétables**. En effet, l'un des points forts de la méthode logistique est l'**interprétabilité** des résultats. Même s'il était possible de renforcer les qualités prédictives de notre modèle en introduisant notamment des variables croisées, nous avons fait le choix (discutable) de privilégier cette force d'interprétation au dépend d'un accroissement (marginal) de la qualité du modèle.

Construction du modèle

01 La méthode **STEPWISE**.

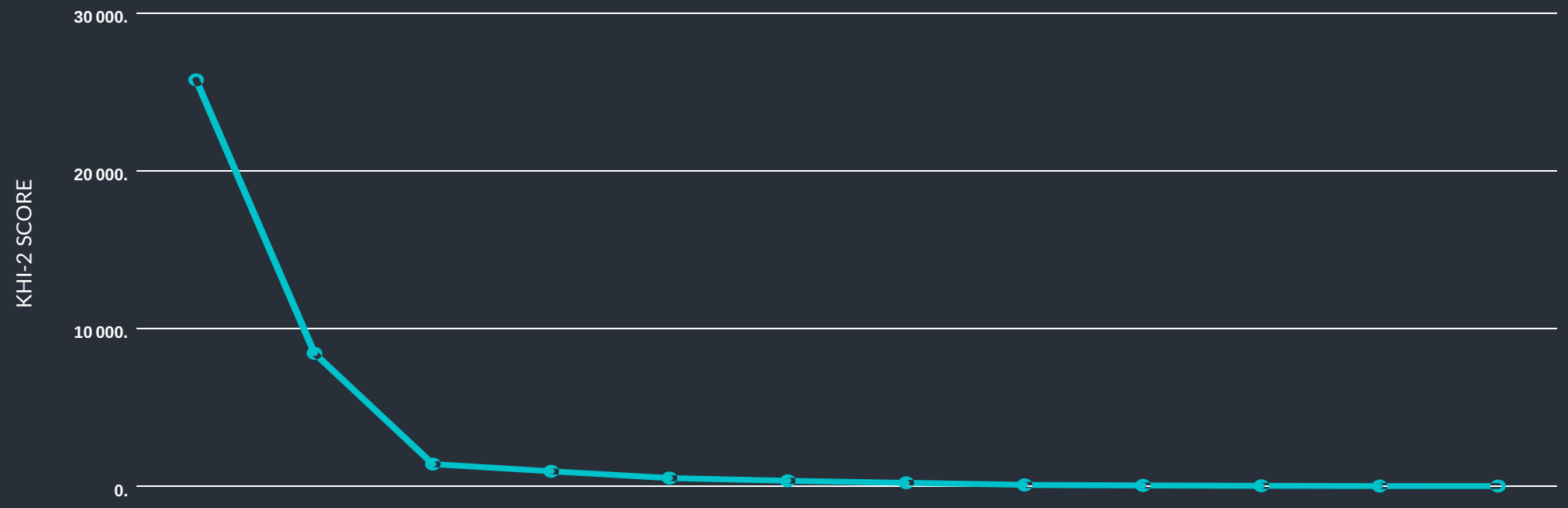
02 La méthode de **l'Adaptive LASSO**.

03 La méthode de **l'Elastic Net**.

La méthode STEPWISE

La méthode **STEPWISE** est une **procédure de sélection** qui permet de réexaminer toutes les variables introduites précédemment dans le modèle. En effet, une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative.

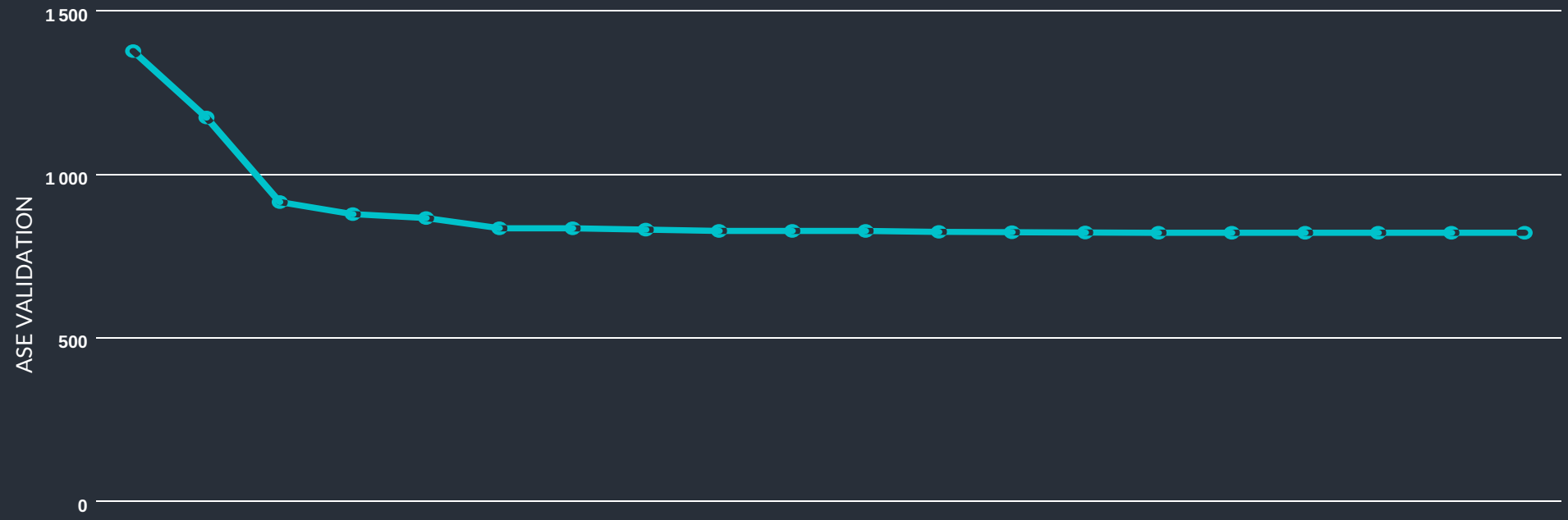
Après l'introduction d'une variable dans le modèle, les **tests de Chi-deux** sont réexaminées pour chaque variable explicative anciennement admise dans le modèle. On retire alors du modèle la **moins significative d'entre elles**.



La méthode de l’Adaptive LASSO

La méthode de **l’Adaptive LASSO** est une méthode de **sélection de variables** dans un contexte de **régression pénalisée**. L’objectif est d’introduire une contrainte sur les paramètres estimés du modèles afin de pénaliser ceux qui ressortent les moins significatifs.

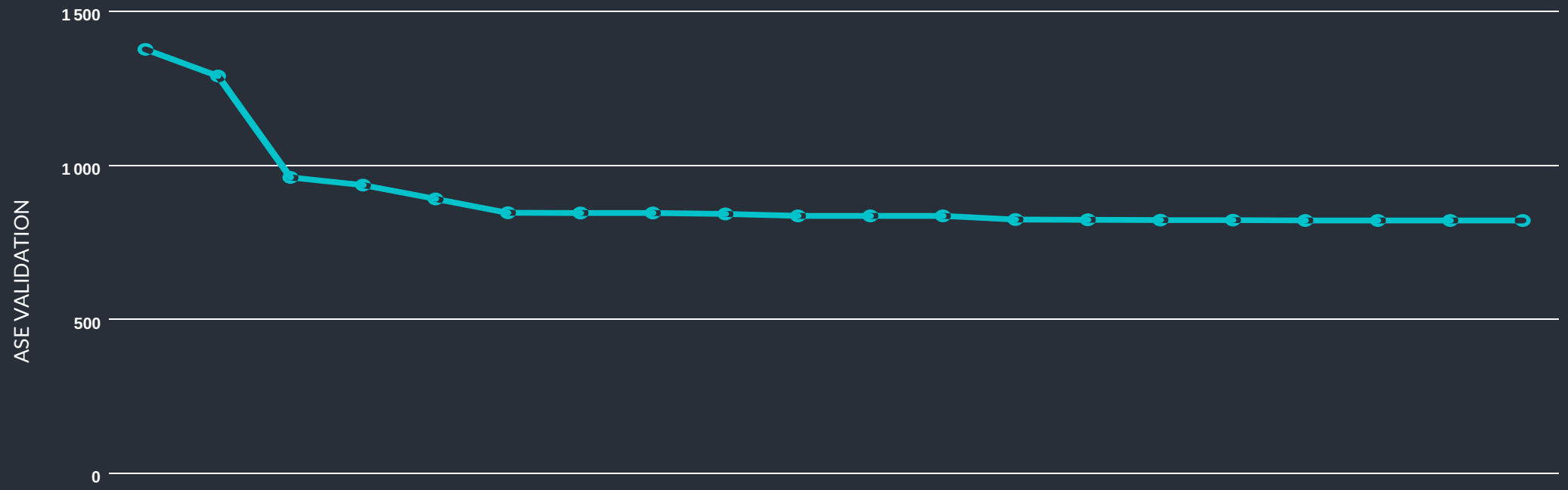
Contrairement à la méthode du LASSO standard, la méthode de l’Adaptive LASSO **force certains coefficients du modèle à être nuls** et les exclus donc du modèle final. En effet, la pénalisation dépend du pouvoir explicatif de chaque variable et n’est plus une simple constante.



La méthode de l’Elastic Net

La méthode de **l’Elastic Net** est une méthode de **sélection de variables** dans un contexte de **régression pénalisée.** Elle combine les **méthodes de pénalisation RIDGE et LASSO**, en introduisant des contraintes de type *norme-1* et *norme-2* sur les coefficients à estimer.

L’Elastic Net est intéressante dans la mesure où elle établie une **sélection de variables groupées**, ce qui permet de palier au risque d’**irreprésentabilité** du LASSO standard.



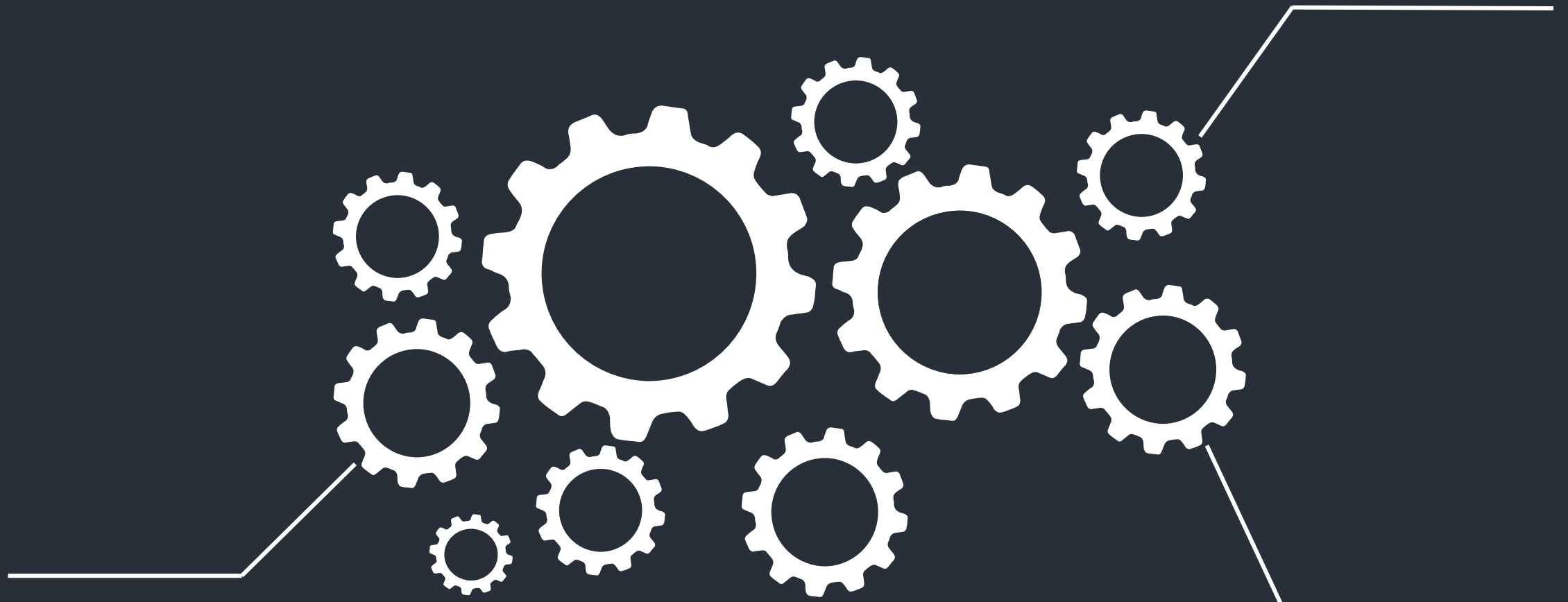
Remarque

Comme nous l'avons vu précédemment, **toutes les variables sont significatives** dans notre modèle. C'est donc au niveau de **l'expertise métier** et de la **parcimonie** que notre choix de variables s'est porté.

D'un point de vue parcimonique, nous avons décidé de garder les variables qui permettent de **diminuer très fortement l'erreur quadratique approchée** (dans le cadre d'une variable dépendante qualitative). Elles étaient sensiblement les mêmes à chaque fois, avec un fort pic au niveau de la 6ème variable.

D'un point de vue métier, nous avons essayé de comprendre **l'enjeu des variables** sélectionnées dans le cadre de la determination d'un indicateur de défaut, et de comparer celles qui nous semblaient plus subtiles à expliquer avec celles qui **ressortaient souvent au niveau métier**. Ce fut par exemple le cas pour le secteur d'activité.

III – Estimations et grille de score



Interprétations

Après avoir estimé notre modèle avec une ***modélisation logistique***, nous pouvons facilement interpréter nos résultats, notamment le ***signe des coefficients estimés***, ***leur rapport de côte*** ainsi que ***les différentes implications*** que vont engendrer notre modèle et notre analyse sur les décisions de l'analyste financier.

[illegible]

Interprétation des signes de nos coefficients :

- Finalement, nous avons donc estimé notre modèle de telle manière que **la modalité mise en référence correspond à chaque fois à la modalité la plus risquée.**

- Cela nous indique donc que notre score, note attribuée à chaque individu en fonction de son niveau de risque, sera une fonction **croissante de sa probabilité de défaut**. Autrement dit, nous nous attendons à ce que plus le score d'un individu soit élevé, plus ce dernier sera risqué.

- Il aurait évidemment été possible de s'établir différemment, en créant un score qui serait une fonction décroissante de la probabilité de défaut.

Interprétation des rapports de côte :

- Le rapport de côte désigne le rapport entre la côte, c'est-à-dire le rapport de chances d'un évènement, issue d'une estimation où notre variable d'intérêt prend la modalité j et la côte issue de l'estimation où notre variable d'intérêt prend la modalité de référence.

- Il peut donc s'interpréter comme un **coefficient de variation** qui permet de mesurer la variation de côte en tenant compte d'une modalité plutôt que de la modalité de référence.

- En ce sens, un **rapport de côte dont la valeur 1 est comprise dans l'intervalle de confiance à 95%** signifie que, pour un risque de 5%, il n'y a pas de changement de côte significatif engendré par les différentes modalités de la variable d'intérêt. Autrement dit, que l'on choisisse une modalité ou une autre de la variable, cela **ne modifie pas le rapport de chances de l'évènement à modéliser**.

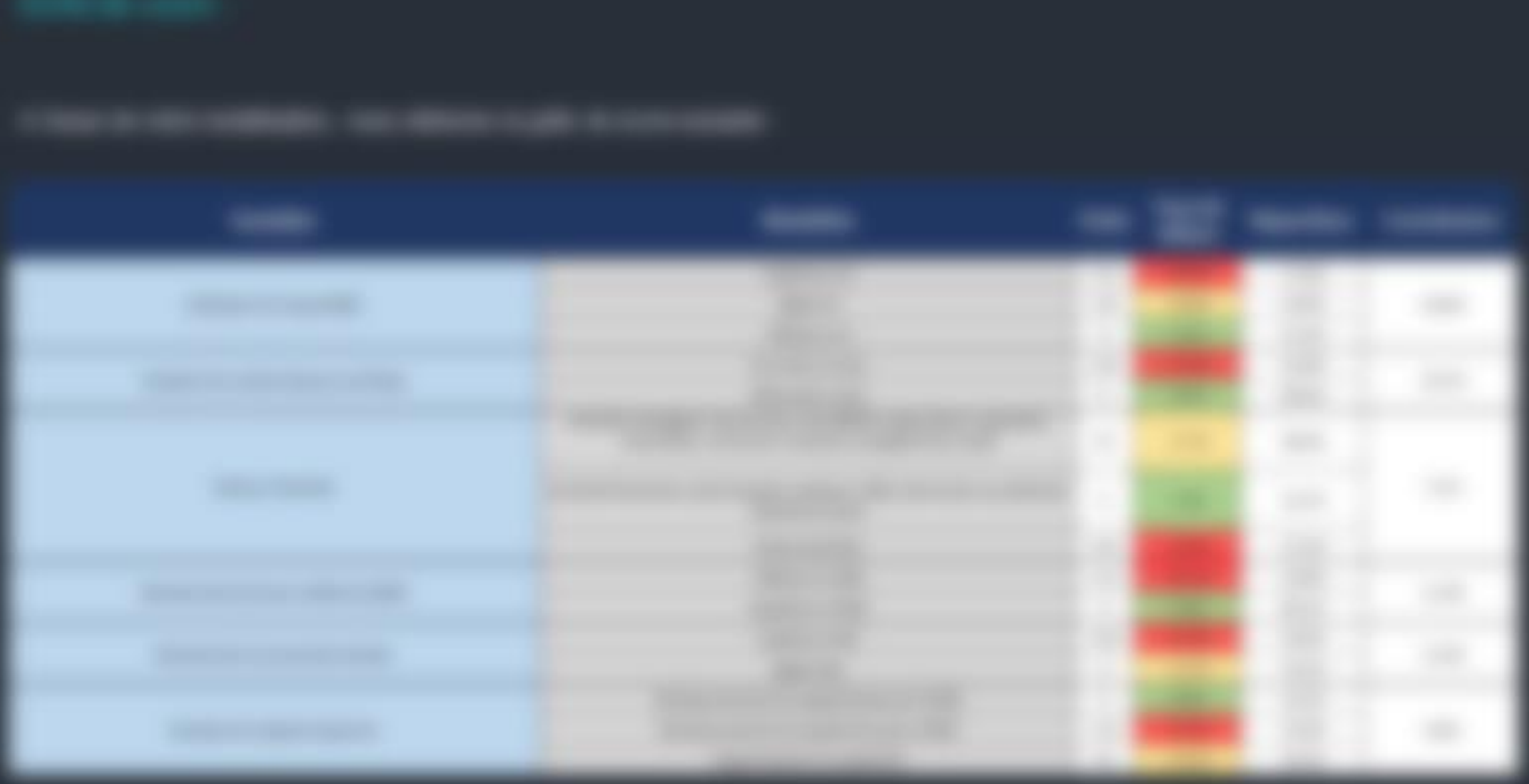
- Par ailleurs, **si notre rapport de côte est supérieur à l'unité**, cela signifie que la modalité choisie permet **d'augmenter le rapport de chances de l'évènement à modéliser**. Autrement dit, **on augmente ses chances de connaître l'évènement par rapport au fait de ne pas connaître l'évènement.**

- Au contraire, **si notre rapport de côte est inférieur à l'unité**, cela signifie que la modalité choisie **diminue le rapport de chances de l'évènement à modéliser**. Autrement dit, on **diminue ses chances de connaître l'évènement par rapport au fait de ne pas connaître l'évènement.**

- Dans notre cas, nous voyons que les rapports de côte complètent l'interprétation précédente sur les signes des coefficients : étant tous inférieurs à 1 et significatifs, cela signifie que **les modalités en référence sont celles qui permettent d'augmenter le rapport de chances et donc de connaître l'évènement.**

Grille de score

Le **scoring (notamment le score de risque)** consiste en **une méthode de segmentation des individus** en fonction de leur **risque respectif** . L'idée est d'attribuer une note à chaque individu et de pouvoir, grâce à cette unique note, construire une aide à la décision compte à l'acceptation ou non d'un octroi de crédit par exemple. Dans ce qui suit, nous avons choisi de **construire un score croissant à la probabilité de faire défaut**, mais l'inverse aurait évidemment été possible.



Grille de score :

Pour rappel, nous avons construit notre grille de score de telle façon que **la note attribuée aux différents individus augmente avec leur probabilité de défaut**. Dès lors, nous pouvons voir que notre grille de score est **cohérente** dans la mesure où:

1. **Les poids sont cohérents avec les taux de défaut associés** : en effet dans notre cas, nous voyons que plus une modalité a de points attribués par notre score, plus la probabilité de défaut sera élevée.

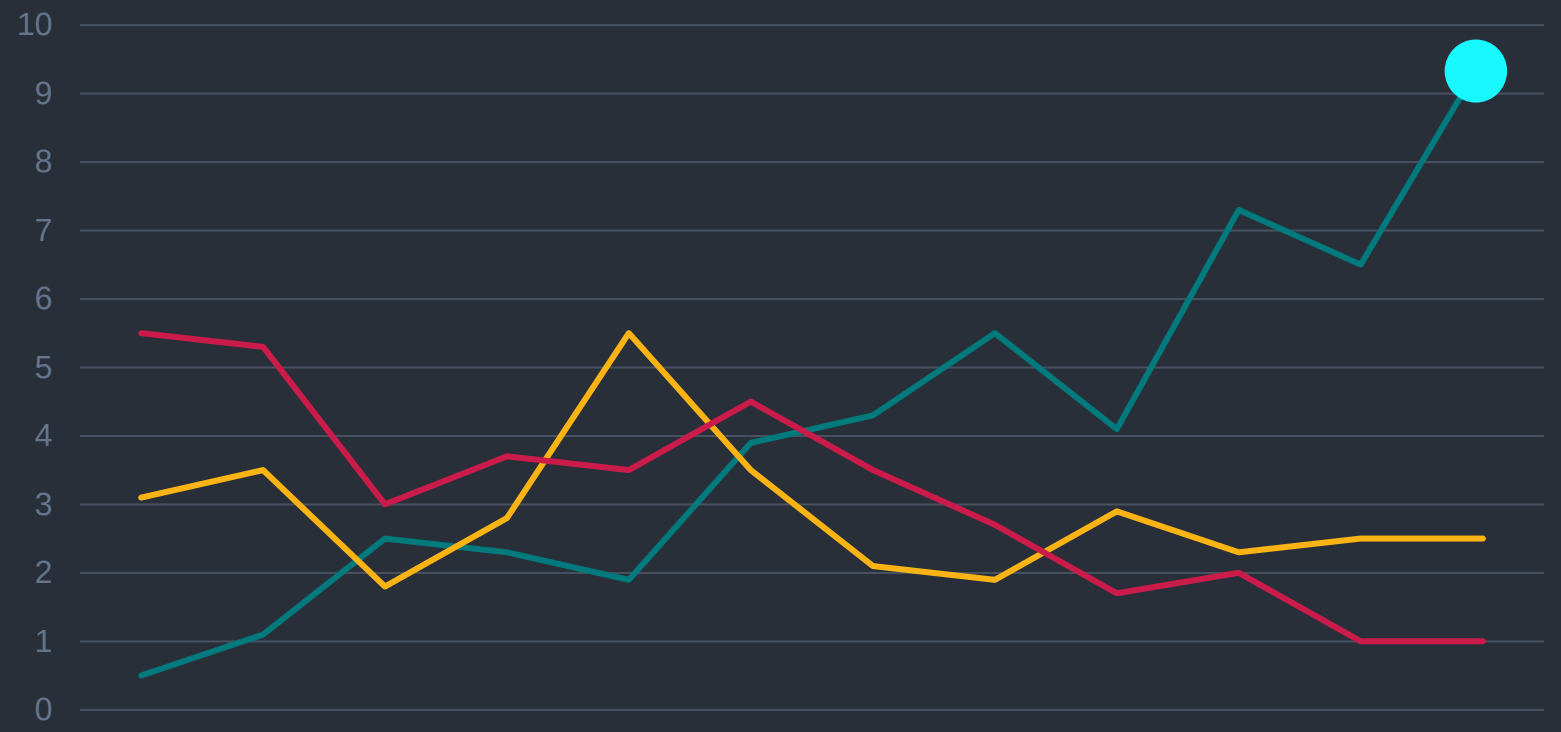
   Autrement dit, **notre score attribue plus de points pour les modalités avec un fort taux de défaut.**

1. Les variables « **Indicateur de risque** » et « **Variation du nombre de jours d'arriérés** » sont les plus contributrices, ce qui, d'un point de vue métier, semble **logique**. En effet, on s'attend à ce que les variables qui permettent le plus d'expliquer un défaut soient **le niveau de risque du contrat** ainsi que la **variation du nombre de jours d'arriérés** (qui est d'ailleurs l'un des critères d'entrée en défaut sous Bâle IV, si on suppose les seuils franchis).

1. Au sein de nos différentes variables, nous pouvons remarquer que **le taux de défaut moyen de notre grille** correspond au **taux de défaut moyen de notre table d'origine**. En effet, nous retrouvons systématiquement un taux de défaut proche de **16.4%** au sein de chacune de nos variables.

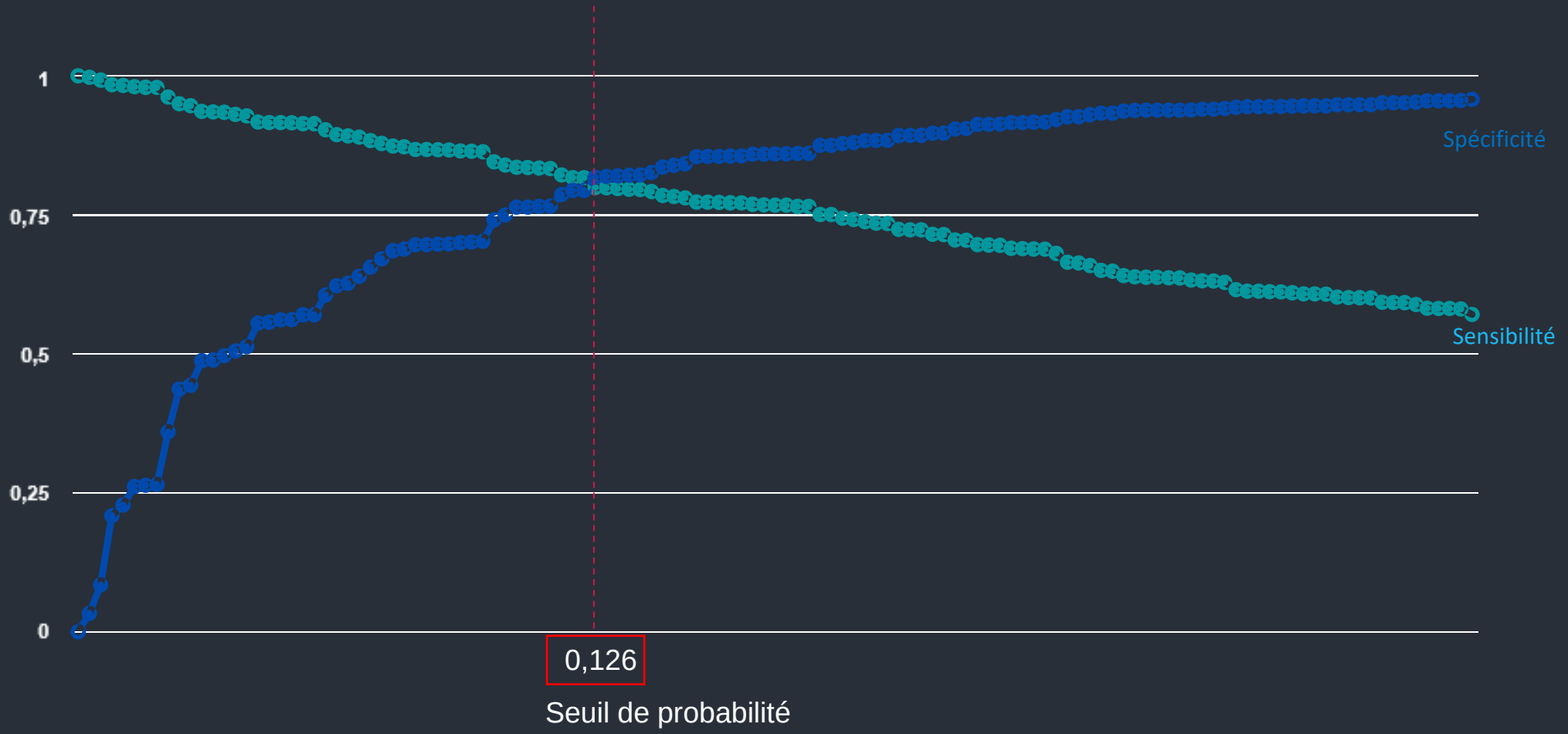
1. Les points maximaux de notre grille de score **se somment à 1000**.

IV – Analyse des performances



Détermination du cut-off optimal pour notre modèle :

D'un point de vue statistique, le **cut-off optimal** s'obtient au **point d'intersection** entre la **sensibilité**, c'est-à-dire la proportion d'individus ayant connu l'évènement et correctement classés par notre score, et la **spécificité**, c'est-à-dire la proportion d'individus n'ayant pas connu l'évènement et correctement classés par notre score.



Détermination du cut-off optimal pour notre modèle :

Ici, nous obtenons un cut-off optimal égal à environ 0,12. Cela signifie que si la probabilité estimée qu'un individu fasse défaut est supérieur à 12%, alors cet individu sera classé comme défaut par notre modèle.

Ce seuil de 12% peut paraître relativement faible, mais n'oublions pas que dans le cadre de notre étude :

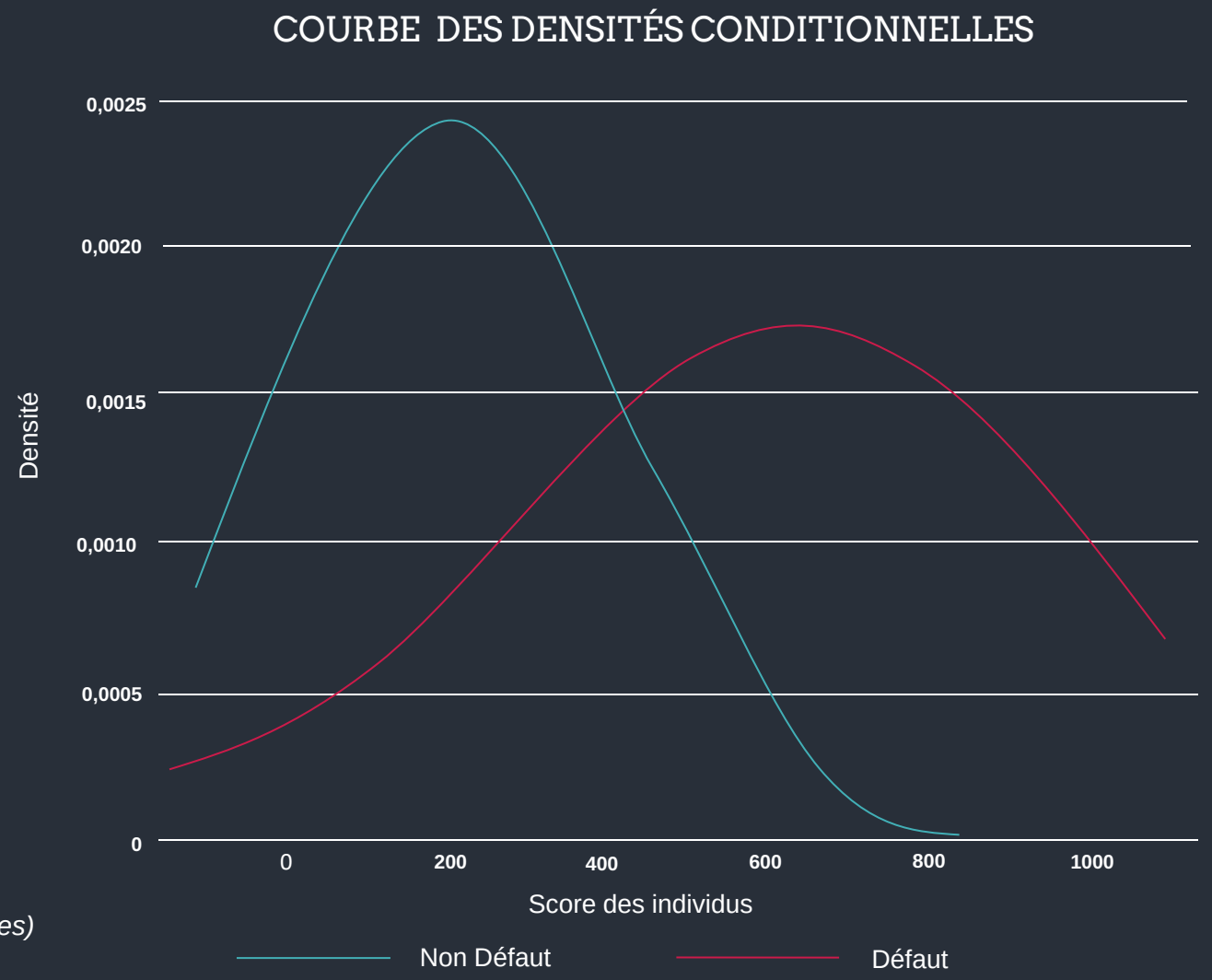
- Nous travaillons sur **une base de clients en anomalie**, c'est-à-dire des clients qui potentiellement, rentrerons en défaut plus facilement que d'autres clients considérés comme « purement sains » par la banque.

- Nous travaillons sur une **perspective de minimiser le risque de type I de la banque**, c'est-à-dire d'attribuer un prêt à un mauvais client, même si cela lui coûte l'opportunité potentielle de refuser un « bon » client.

- Nous ne disposons pas de la **part de marché de la banque**, ni de ces **paramètres internes**, et ne pouvons donc pas apporter davantage de précision compte à ce seuil, optimal du point de vue statistique.

Statistiques de performance :

Avant d'analyser les performances de notre modèle logistique, nous avons représenté **la distribution des scores des individus de notre base de donnée, conditionnellement au défaut**. Plus les distributions sont éloignées, plus le score est discriminant.



Graphique version SAS

Graphique version SAS (avec histogrammes)

Statistiques de performance :

a. Statistiques liées au nombre de paires concordantes :

Soient deux individus de notre base notés i et j. Supposons que l'individu i ait connu l'évènement, alors que l'individu j n'ait pas connu l'évènement. Ces individus constituent **une paire concordante** si **la probabilité que l'individu i connaisse l'évènement est supérieure à la probabilité que l'individu j connaisse l'évènement.**

De cette définition émergent différentes mesures de performance, notamment **la statistique c (ou l'AUC)** et le **coefficient de Gini**. Pour ces statistiques, plus leur valeur sera élevée (bornée à 1), meilleure sera la qualité prédictive du modèle.

Pourcentage de paires concordantes	AUC	Coefficient de Gini
0,879	0,884	0,767

Notre **AUC** nous indique que si on considère deux individus au hasard dans notre base, l'un ayant fait défaut et l'autre n'ayant pas fait défaut, alors notre modèle attribuera une probabilité de défaut plus élevée à l' individu ayant réellement fait défaut **dans 88,4% des cas**.

Statistiques de performance :

b. Statistiques liées à la matrice de confusion :

La **matrice de confusion** est un outil permettant de comparer les résultats issus de la prédiction de notre modèle avec les résultats observés de notre échantillon test.

De cette outil émergent différentes mesures de performances, notamment l'**accuracy,** la **sensibilité**, **la spécificité**, **le rapport de vraisemblance positif et le rapport de vraisemblance négatif**.

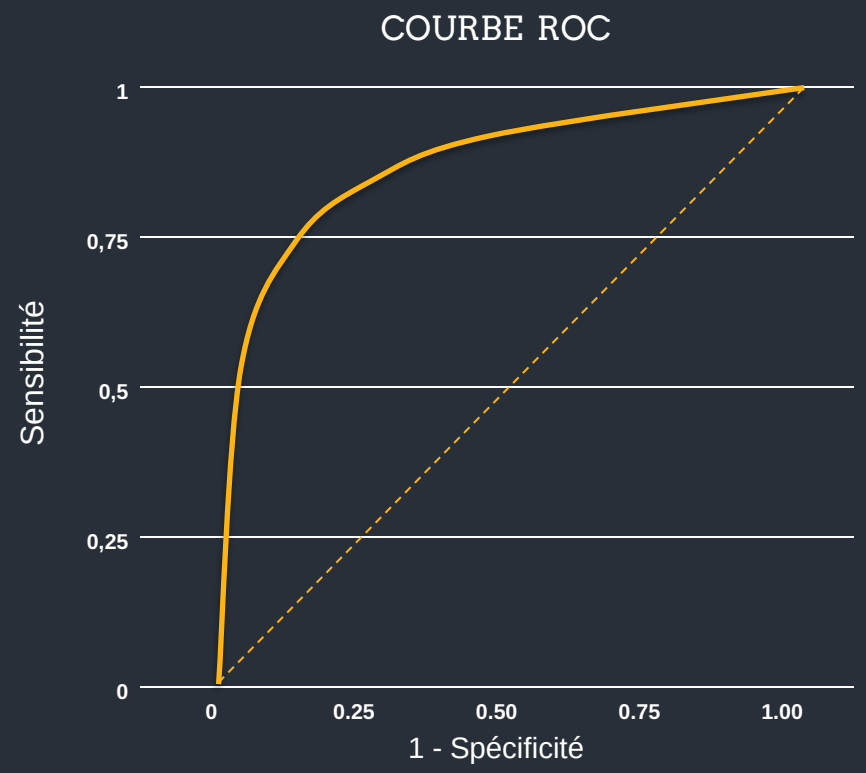
Seuil utilisé	Accuracy	Sensibilité	Spécificité	Rapport de vraisemblance positif	Rapport de vraisemblance négatif
0,126	0,817	0,800	0,821	4,447	0,248

Notre modèle permet **d'identifier correctement 82% des individus n'ayant pas fait défaut** et **80% des individus ayant fait défaut**. Par ailleurs**, il classe correctement un individu avec une probabilité de 81,7%**. Enfin, un individu **a 4,45 fois plus de chance de faire défaut en réalité s'il a été identifié comme tel par le modèle** et au contraire, un individu a environ 0,25 fois plus de chances, c'est-à-dire **4 fois moins de chances, d'être en défaut dans la vraie vie s'il a été identifié comme en non défaut par le modèle.**

Statistiques de performance :

c. Analyse graphique des performances :

Il existe différents graphiques pour mesurer les performances d'un modèle logistique. Nous allons en présenter deux, à savoir la **courbe ROC** et la **courbe de sélection**.



Graphique version SAS

Plus la courbe ROC se rapproche du coin externe gauche en hauteur, meilleur le modèle sera.

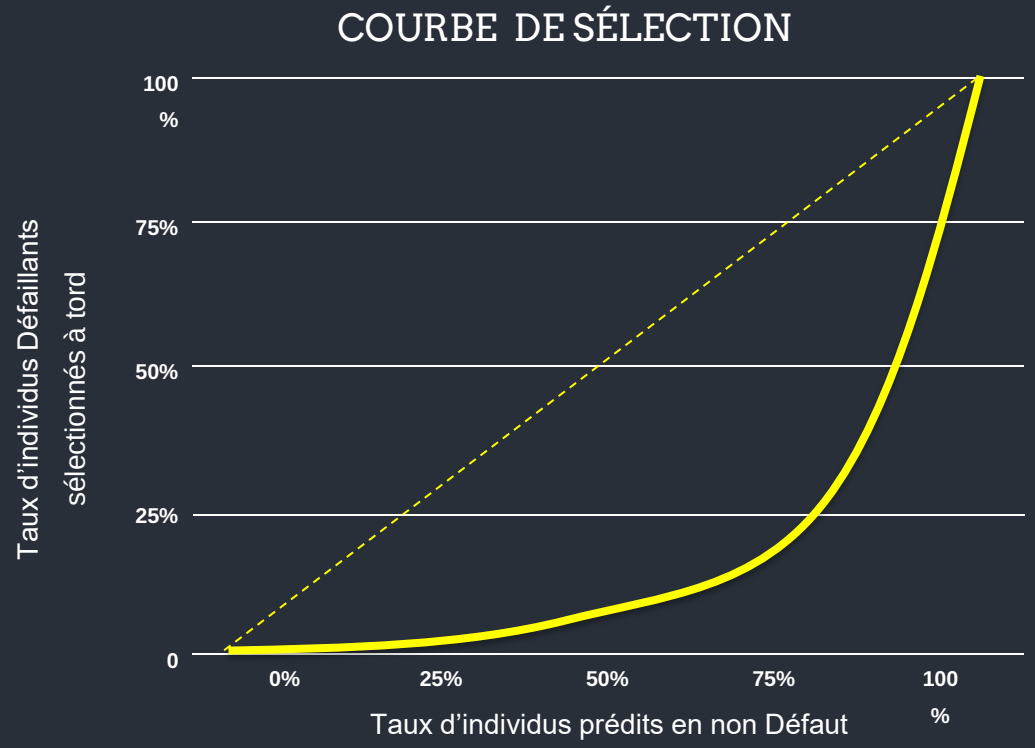
Au contraire, si la courbe ROC se rapproche de la première bissectrice, notre modèle ne sera pas plus performant qu'une classification aléatoire.

Statistiques de performance :

c. Analyse graphique des performances :

La **courbe de selection** est un outil graphique permettant de mesurer **le taux d'individus défaillants sélectionnés à tord par notre modèle en fonction de la part de marché de la banque**, c'est-à-dire du taux d'individus prédits comme bon payeurs par notre score.

La part de marché est un paramètre interne à la banque. Cette courbe peut donc lui permettre de determiner son cut-off optimal, conditionnellement à ses caractéristiques.



Graphique version SAS

V – Création de classes de risque

Exemple

Le point de [redacted] est [redacted], nous pouvons classer nos clients en différentes catégories de risque. Voici nos [redacted] :

Classe	Score attribué	Probabilité de défaut	Risque client
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

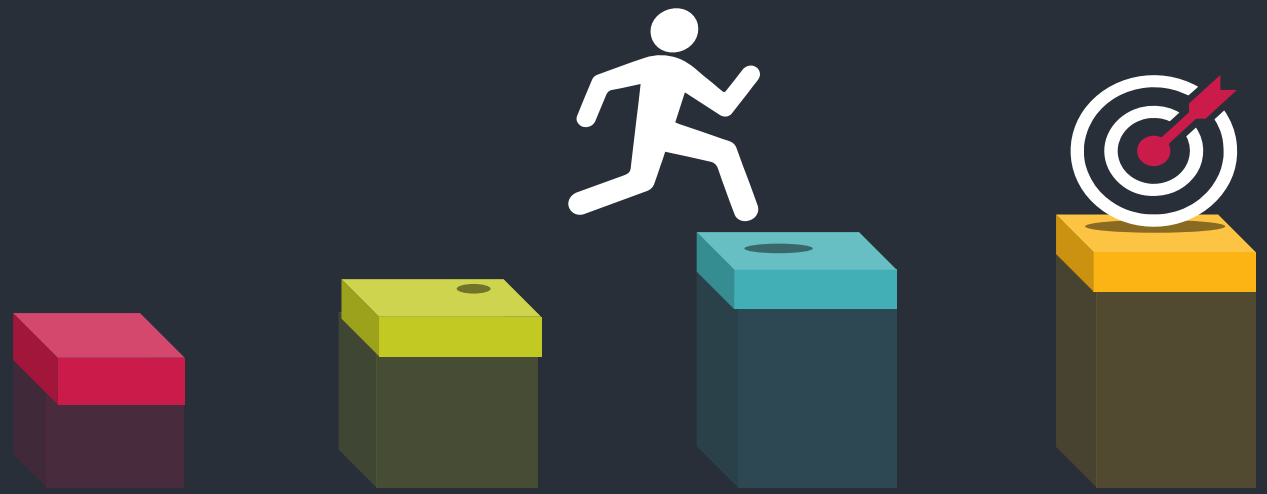
Conclusion de la première partie

- La création d'une grille de score est une étape fondamentale de l'industrie bancaire. Il est donc nécessaire que les banques soient capables de présenter **des données de qualité sur des bases de risques.**

- Les scores attribués aux différents clients peuvent impacter **le calcul des risques pour les fonds propres**. Ainsi, si la banque réussit au mieux à catégoriser ses clients, le **calcul des RWA pourra être amélioré**. Au contraire, si la banque se trompe, alors les décisions prises par rapport aux clients peuvent impacter fortement les données comptables de la banque, comme basculer brutalement le compte de résultat dans le négatif.

- Nous avons aboutis à la construction de **5 classes de risques sous SAS**, mais il était également possible d'utiliser une méthode de clustering telle que **les kmeans (**que nous avons lancée sous R) qui a abouti à un nombre optimal de classes égal à 4.

- Le modèle Logit nous a globalement fourni de **bons résultats** qui ont l'avantage d'être **facilement interprétables**. D'autres méthodes de Machine Learning pourront peut-être fournir de meilleurs résultats, au détriment d'une explicabilité plus souple.



Résultat R

Sources première partie

Sources internes :

- *Cours de Scoring - Mr.RAULT*
- *Cours de Machine Learning - Mr.Tokpavi*
- *Cours de Support Vector Machine - Mr.Hurlin*
- *Cours de Classification – Mr. Lahiani*
- *Cours de Règlementation Bancaire – Mr. Hurlin*

Sources externes :

- *https://github.com/xieliaing/SAS/blob/master/ModelValidation/AUC.sas*
- *https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-detaillee-de-sensibilite-et-specificite*

II – Machine Learning et Explicatibilité

- Lors de cette dernière partie, nous allons challenger notre modèle classique logistique avec des modèles plus avancés de Machine Learning. Nous développerons pour cela 5 méthodes de Machine Learning supervisées, à savoir les K-Nearest Neighbors, le Random Forest, le Gradient Boosting, le XGBoost et le SVM.

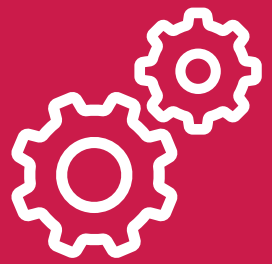
- Lors de l'exécution des 4 premières méthodes, nous avons repris l'intégralité de nos variables, dans la mesure où ces 4 algorithmes sont capables de capter des effets fortement non linéaires et donc d'améliorer fortement les capacités du modèle.

- En revanche, pour des raisons qui seront abordées ultérieurement, nous avons gardé uniquement les variables présélectionnées dans la régression logistique pour lancer le SVM.

- La question ici sera de savoir si un modèle logistique simple et interprétable sera meilleur qu'un modèle plus complexe de machine learning, nécessitant des techniques d'interprétation plus avancées.

METHODES

de Machine Learning



KNN

Random Forest

Gradient Boosting

XGBoost

SVM

K-Nearest-Neighbors

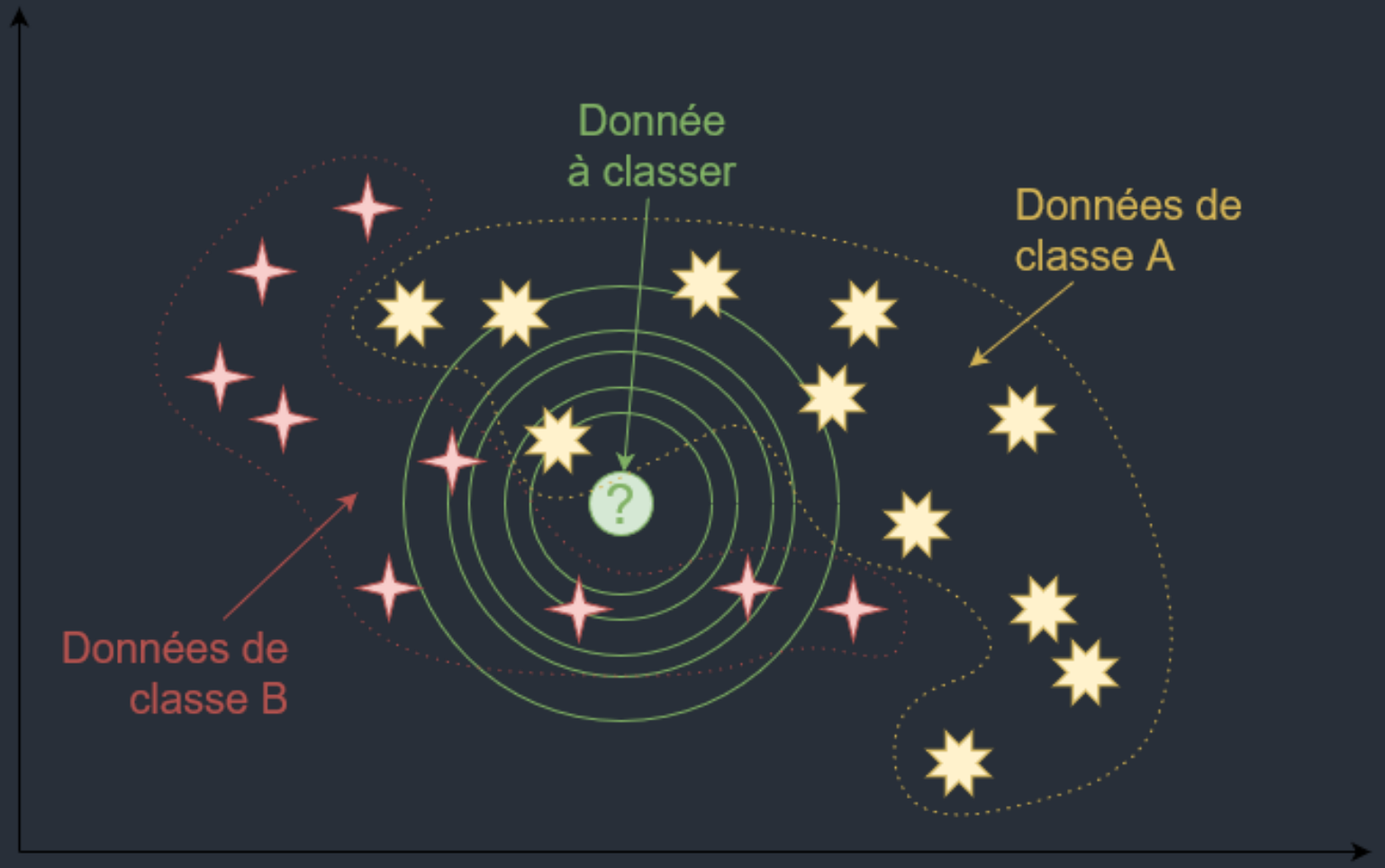
Méthode I – K-Nearest-Neighbors

La méthode des KNN est une méthode **d'algorithme supervisé** simple. Cette méthode a pour but de classifier des points cibles de classe méconnue en fonction de leur distance par rapport aux K points les plus proches, constituant un échantillon d'apprentissage dont la classe est connue à priori.

Cet algorithme s'établie globalement en 5 étapes :

- Sélectionner le nombre K de voisins
- Calculer la distance correspondante
- Prendre les K voisins les plus proches des points non classifiés aux autres points
- Parmi ces K voisins, compter le nombre de points appartenant à chaque catégorie
- Attribuer le nouveau point à la catégorie la plus présente parmi ces K voisins

Méthode I – K-Nearest-Neighbors



Algorithme des KNN

Méthode I – K-Nearest-Neighbors

1) Matrice de Confusion

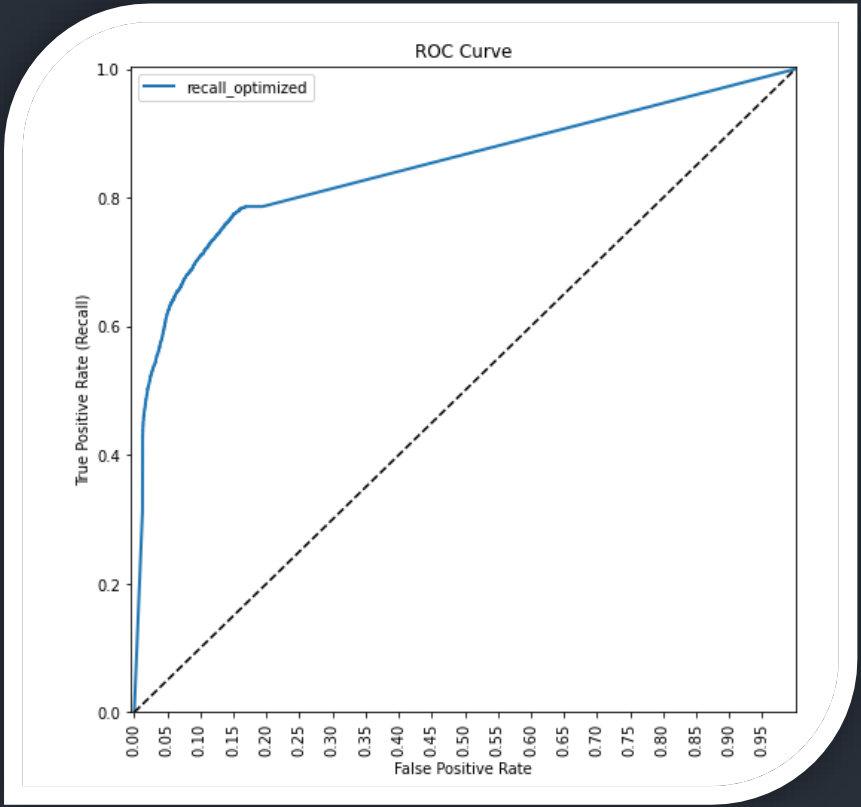
	0 prédits	1 prédits
0 observés	25357	5096
1 observés	1291	4724

Sensibilité : 78,53 %

Spécificité : 53,26 %

Méthode I – K-Nearest-Neighbors

2) Courbe ROC



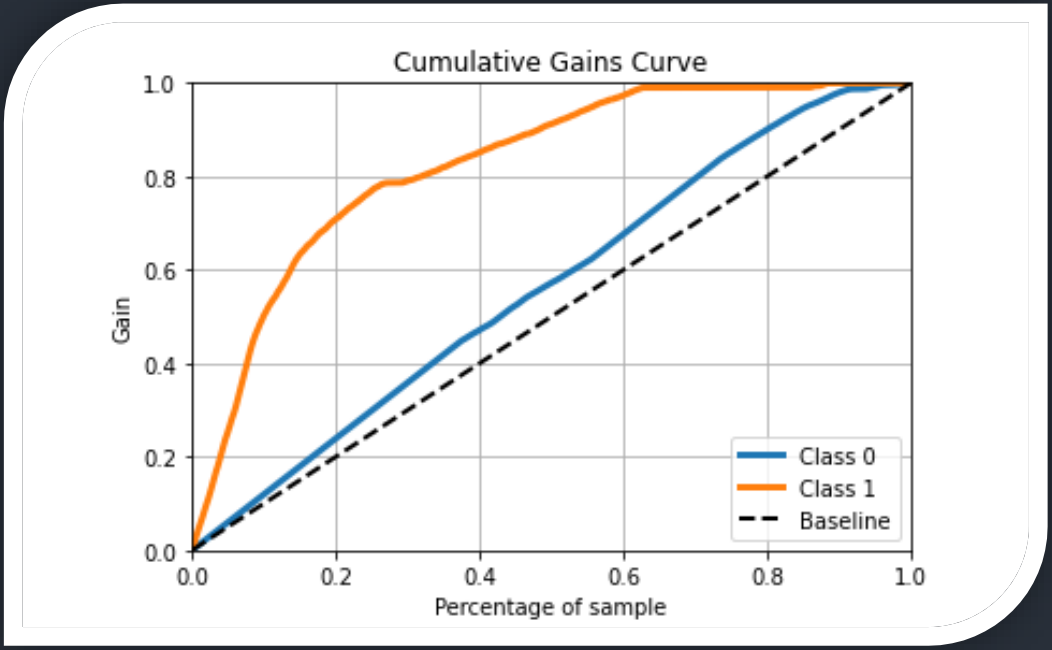
Area-Under-the-Curve : 0,85

Indice de Gini : 0,69

Si l'on choisit deux clients au hasard, un défaillant et un sain, il y a 85% de chance pour que la probabilité de défaut soit plus élevée pour l'individu réellement en défaut.

Méthode I – K-Nearest-Neighbors

3) Courbe des gains cumulés

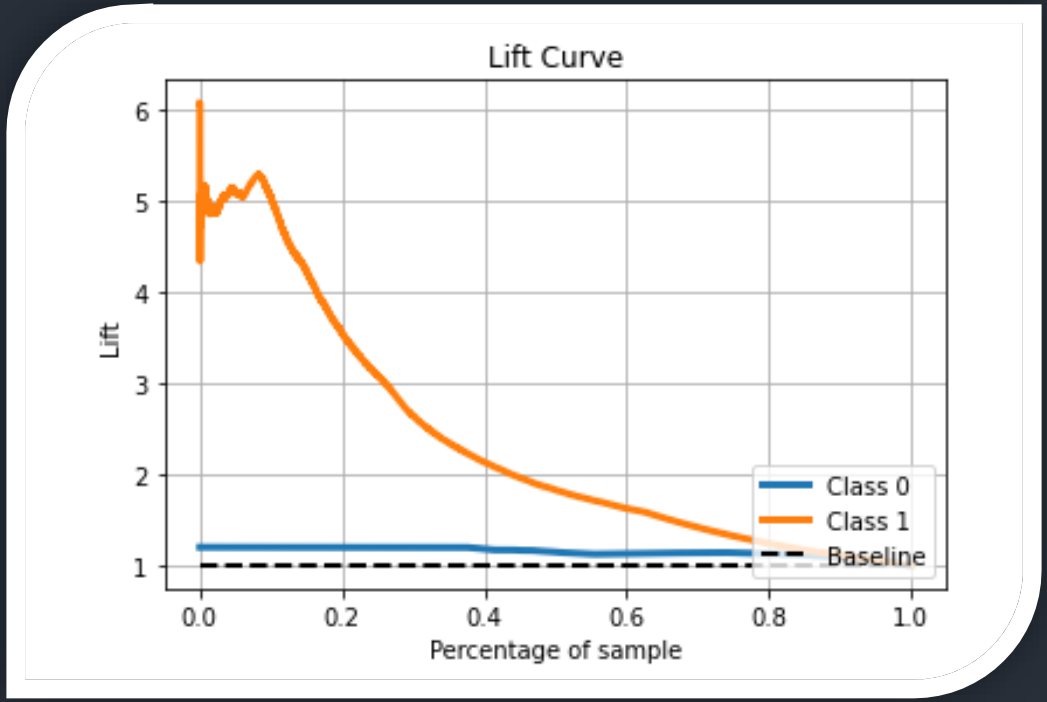


La ligne de référence pointillée représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

Dans notre cas, la courbe des gains augmente fortement au-dessus de la ligne de référence, puis ralentit. Ici, nous voyons que 20% des individus de la base associés aux probabilités de défaut les plus élevées contiendraient environ 70% des clients défaillants

Méthode I – K-Nearest-Neighbors

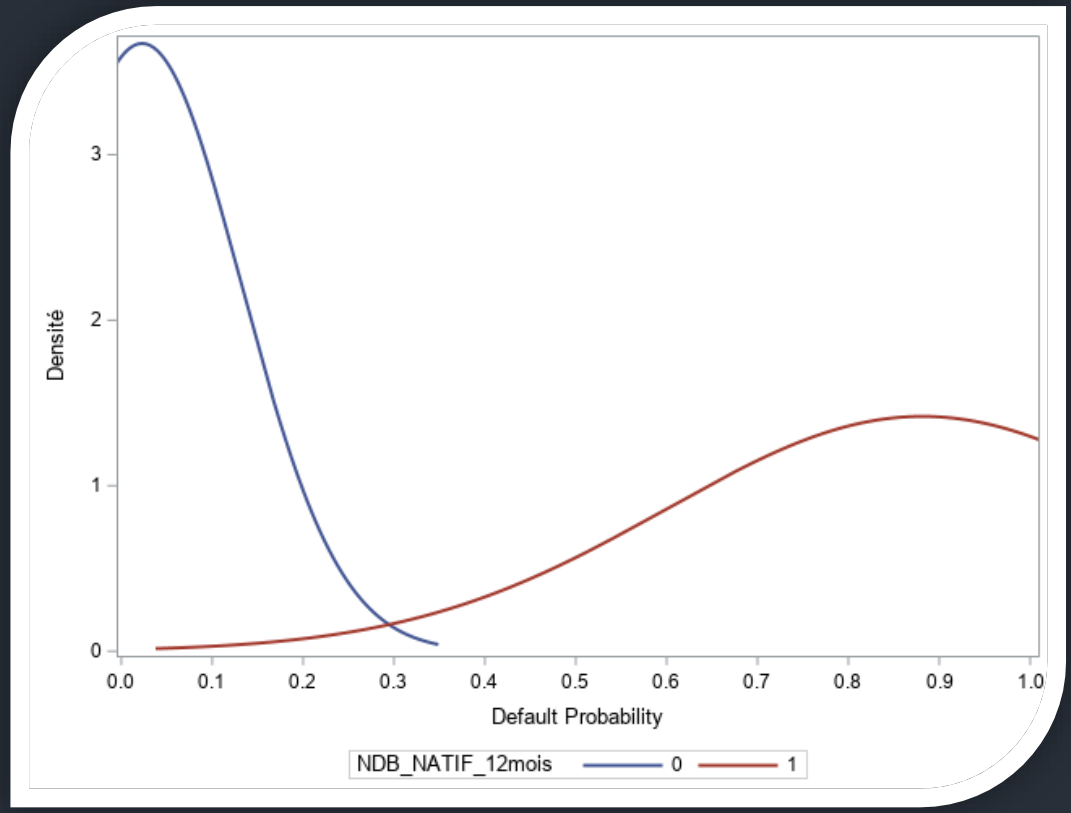
4) Courbe LIFT



La courbe Lift est une dérivée de la courbe de gains cumulés. Ainsi, le Lift à 10% pour les défauts est de 5 environ : en utilisant notre modèle pour prédire les défauts en retenant les 10% des clients avec la probabilité de défaut la plus élevée, nous nous attendons à retrouver 5 fois plus de défauts que si l'on effectuait une requête aléatoire.

Méthode I – K-Nearest Neighbors

Classe de risque	Probabilité de défaut
Peu risquée	[0,0.1634]
Assez risquée	[0.1634,0.498]
Risquée	[0.498,0.8341]
Très risquée	[0.8341,1]



Densités conditionnelles KNN

Le modèle semble très discriminant car nos deux courbes de densité sont très éloignées l'une de l'autre. Cela peut toute fois compliquer la création de plusieurs classes de risque.

Méthode I – K-Nearest-Neighbors

Avantages de la méthode :

- Cet algorithme est simple et facile à mettre en œuvre
- Il n’est pas nécessaire de créer un modèle, ou de régler plusieurs paramètres autre que le nombre de voisins les plus proches

Inconvénient de la méthode :

- L’algorithme devient beaucoup plus lent à mesure que le nombre d’observations et de variables indépendantes augmentent

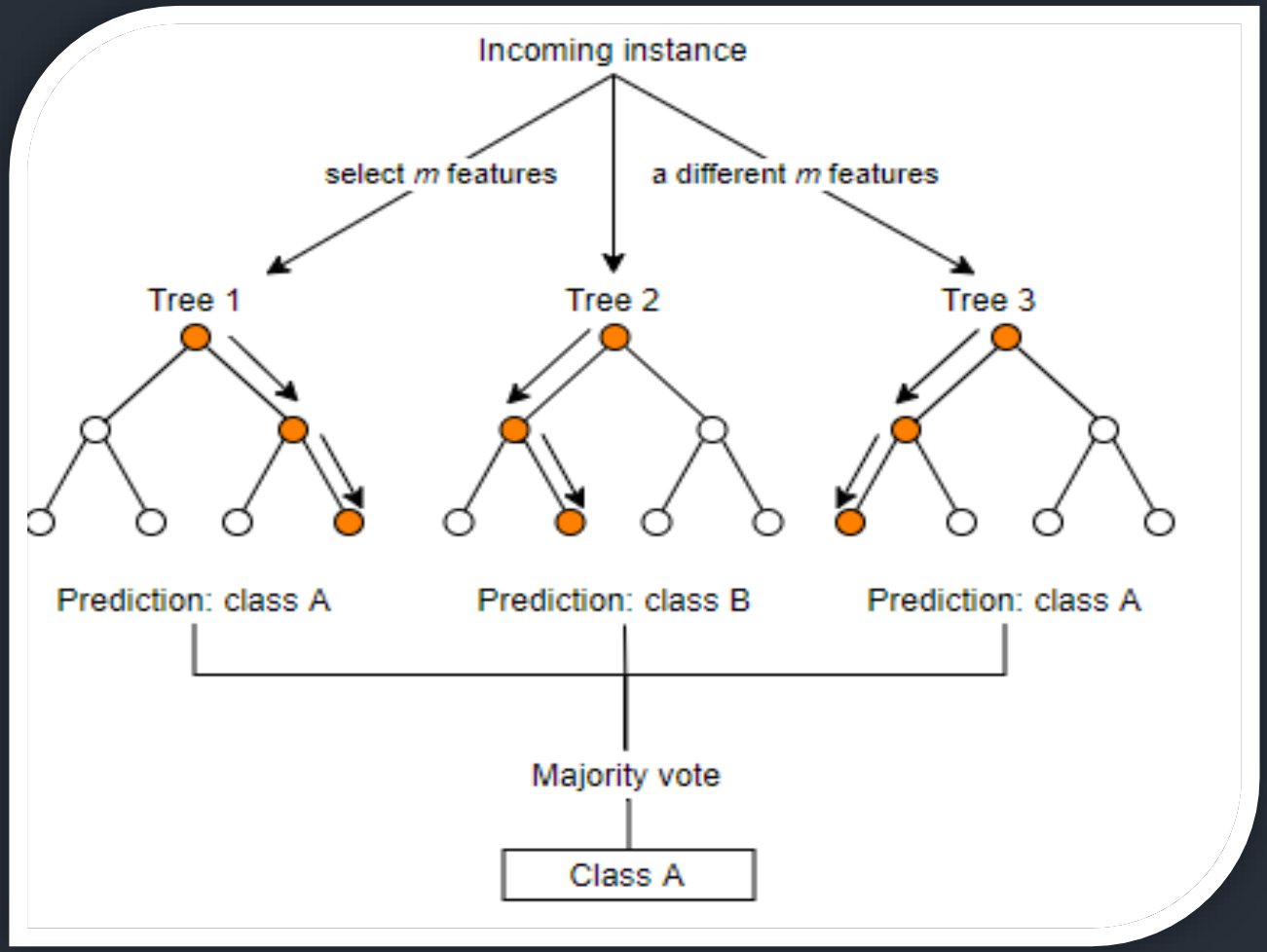
Random Forest

Méthode II – RANDOM FOREST

L’algorithme des forêts aléatoires est un **algorithme de classification** qui réduit **la variance des prévisions d’un arbre de décision seul**, améliorant ainsi leurs performances. Pour cela, il combine de nombreux arbres de décisions dans une approche de type **bagging**.

L’apprentissage s’effectue **en parallèle** sur de multiples arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents. Le nombre idéal d’arbres est un paramètre important et variable en fonction du problème. Un autre paramètre important est la profondeur maximale de chaque arbre, qui constitue un arbitrage entre temps de calcul et sous-apprentissage.

Méthode II – RANDOM FOREST



Algorithme du Random Forest

Méthode II – RANDOM FOREST

1) Matrice de Confusion

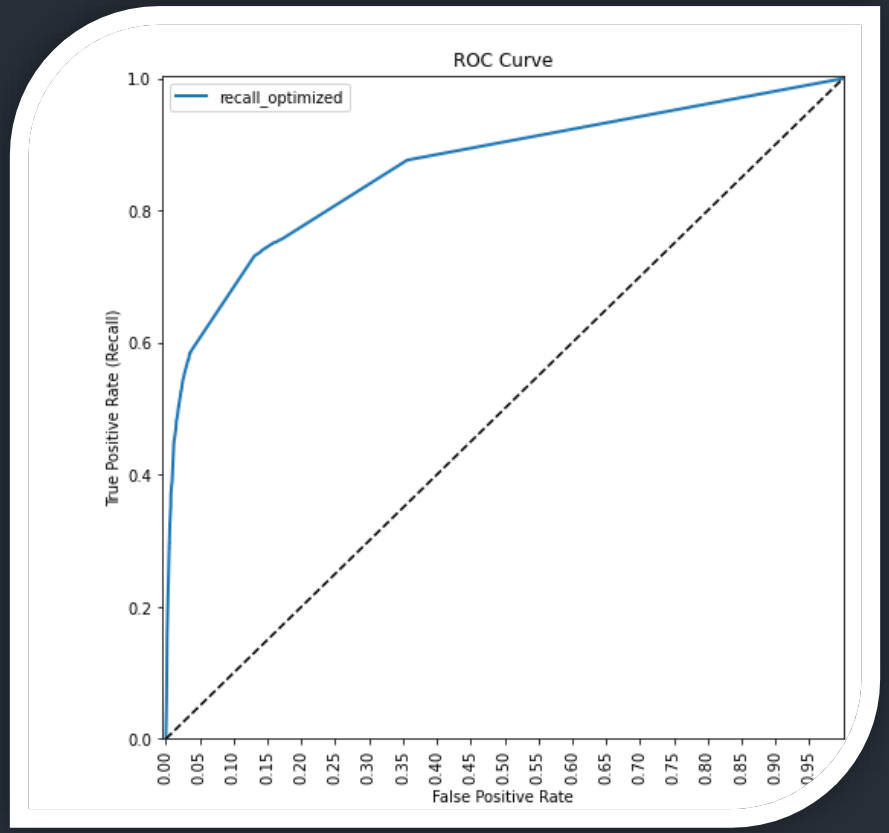
	0 prédits	1 prédits
0 observés	26034	4419
1 observés	1551	4464

Sensibilité : 85,49 %

Spécificité : 74,21 %

Méthode II – RANDOM FOREST

2) Courbe ROC



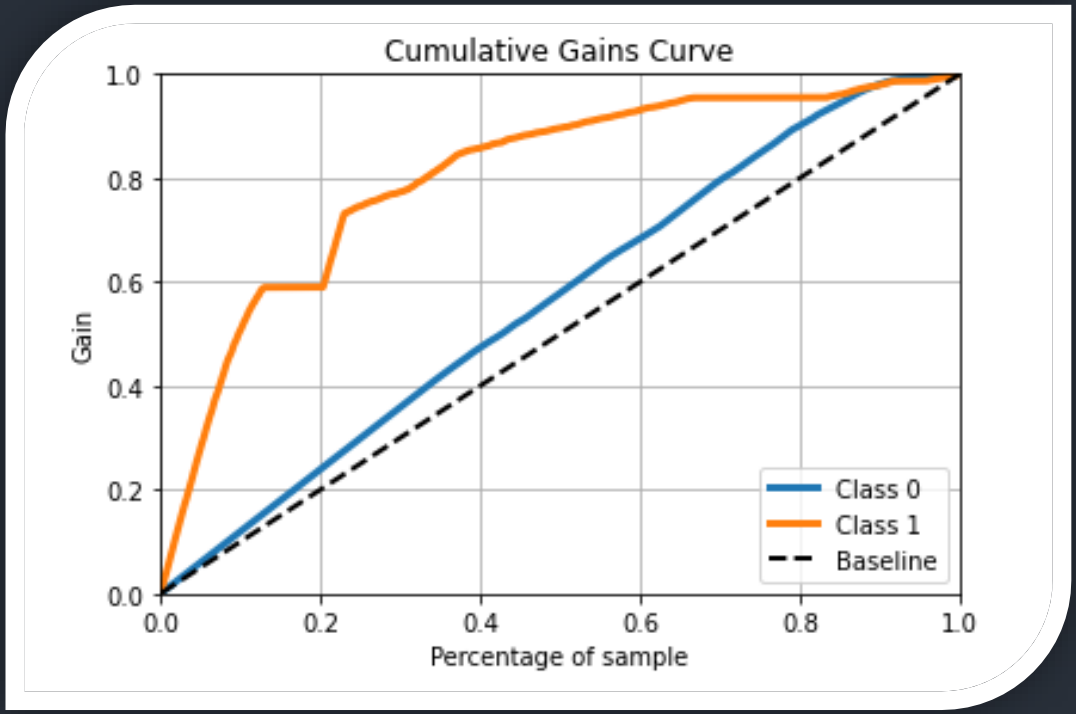
Area-Under-the-Curve : 0,86

Indice de Gini : 0,73

Si l'on choisit deux clients au hasard, un défaillant et un sain, il y a 86% de chance pour que la probabilité de défaut soit plus élevée pour l'individu réellement en défaut.

Méthode II – RANDOM FOREST

3) Courbe des gains cumulés

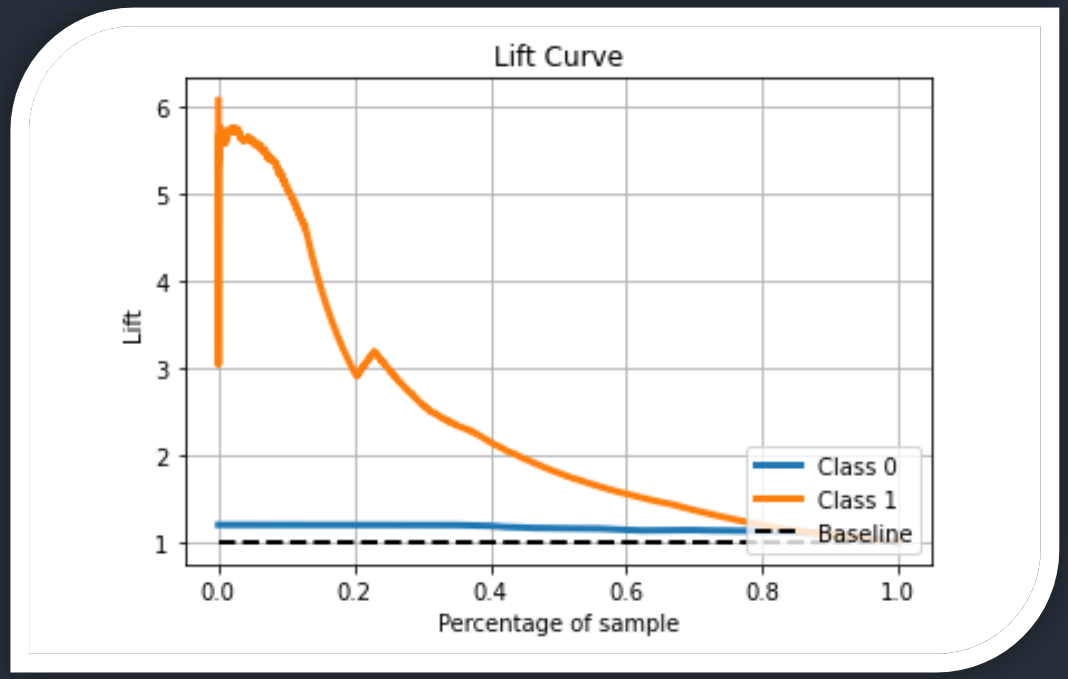


La ligne de référence pointillée représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

Dans notre cas, la courbe des gains augmente fortement au-dessus de la ligne de référence, puis ralentit. Ici, nous voyons que 20% des individus de la base associés aux probabilités de défaut les plus élevées contiendraient environ 60% des clients défaillants

Méthode II – RANDOM FOREST

4) Courbe LIFT

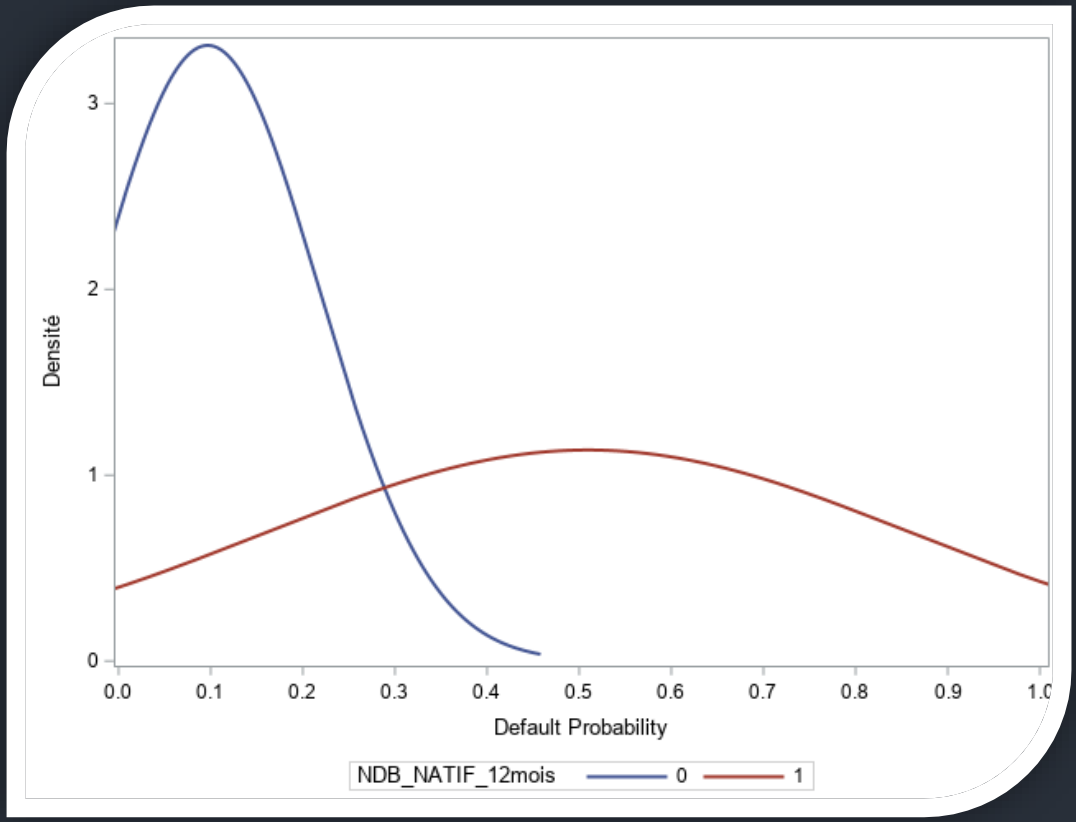


Le Lift à 10% pour les défauts est de 5,5 environ : en utilisant notre modèle pour prédire les défauts en retenant les 10% des clients avec la probabilité de défaut la plus élevée, nous nous attendons à retrouver 5,5 fois plus de défauts que si l'on effectuait une requête aléatoire.

MÉTHODE II – RANDOM FOREST

Classe de risque	Probabilité de défaut
Peu risquée	[0,0.26]
Assez risquée	[0.2575,0.43]
Risquée	[0.43,0.71]
Très risquée	[0.71,0.847]
Quasi en défaut	[0.847,1]

Remarque : Les densités conditionnelles trop éloignées ont compliqué la mise en place de classes de risque.



Densités conditionnelles Random Forest

Le modèle semble très discriminant car nos deux courbes de densité sont bien éloignées l'une de l'autre.

Méthode II – RANDOM FOREST

Avantages du Random Forest :

- Il n’existe pas de risque de sur-apprentissage grâce à la génération naturelle de l’échantillon Out-Of-Sample
- Les paramètres sont faciles à calibrer, c’est pourquoi cet algorithme est souvent utilisé comme benchmark dans les compétitions de Machine Learning

Inconvénients du Random Forest :

- Cet algorithme est difficilement améliorable et est lent à entraîner.

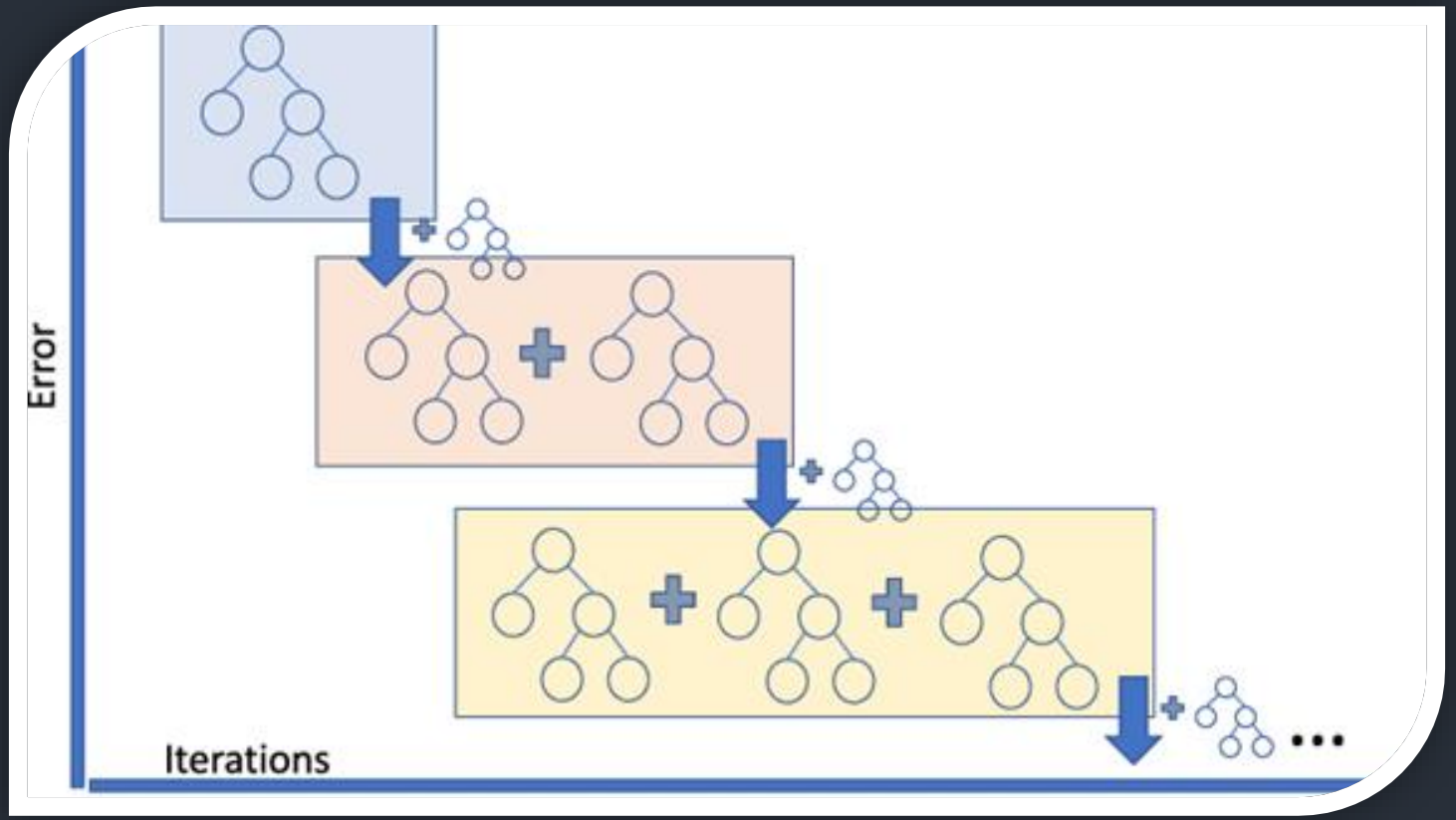
Gradient Boosting

Méthode III – GRADIENT BOOSTING

La **méthode du Boosting** est une méthode permettant de transformer les apprenants faibles en apprenants forts. Dans le paysage du boosting, chaque arbre s'adapte à la version modifiée du premier ensemble de données.

Le Gradient Boosting est un type de boosting. Il repose fortement sur la prédiction que le prochain modèle réduira les erreurs de prédiction lorsqu'il sera mélangé avec les précédents. Cette approche propose de cherche la meilleure combinaison linéaire d'arbres binaires, en utilisant **une descente de gradient.**

Méthode III – GRADIENT BOOSTING



Algorithme du Gradient Boosting

Méthode III – GRADIENT BOOSTING

1) Matrice de Confusion

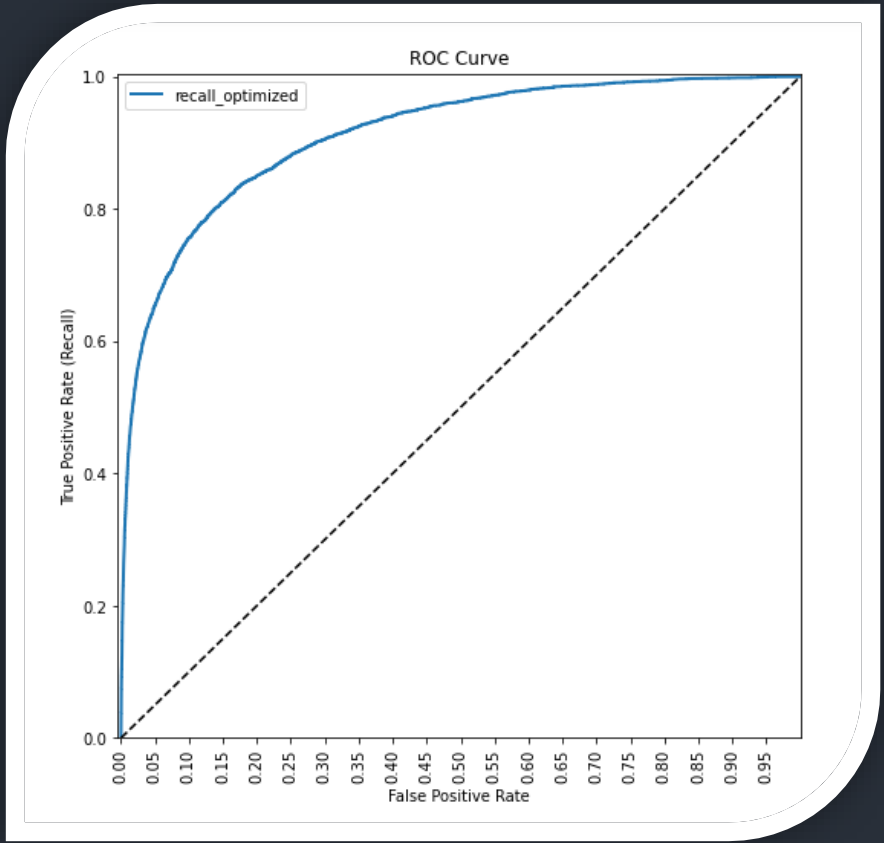
	0 prédits	1 prédits
0 observés	26748	3705
1 observés	1319	4696

Sensibilité : 87,83 %

Spécificité : 78,07 %

Méthode III – GRADIENT BOOSTING

2) Courbe ROC



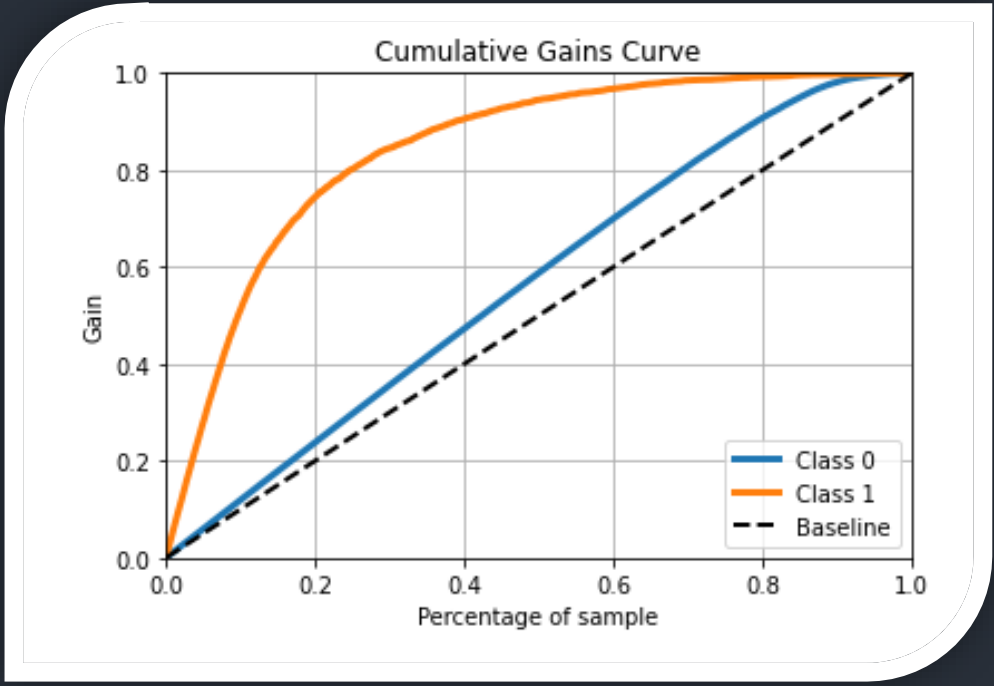
Area-Under-the-Curve : 0,91

Indice de Gini : 0,82

Si l'on choisit deux clients au hasard, un défaillant et un sain, il y a 91% de chance pour que la probabilité de défaut soit plus élevée pour l'individu réellement en défaut.

Méthode III – GRADIENT BOOSTING

3) Courbe des gains cumulés

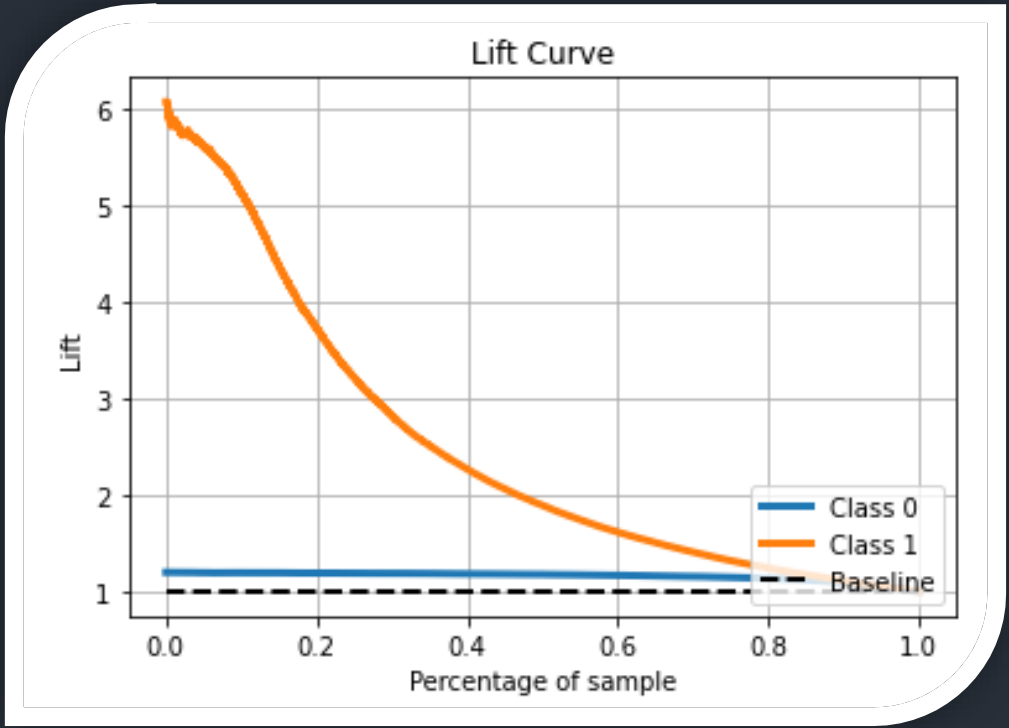


La ligne de référence pointillée représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

Ici, nous voyons que 20% des individus de la base associés aux probabilités de défaut les plus élevées contiendraient environ 75% des clients défaillants.

Méthode III – GRADIENT BOOSTING

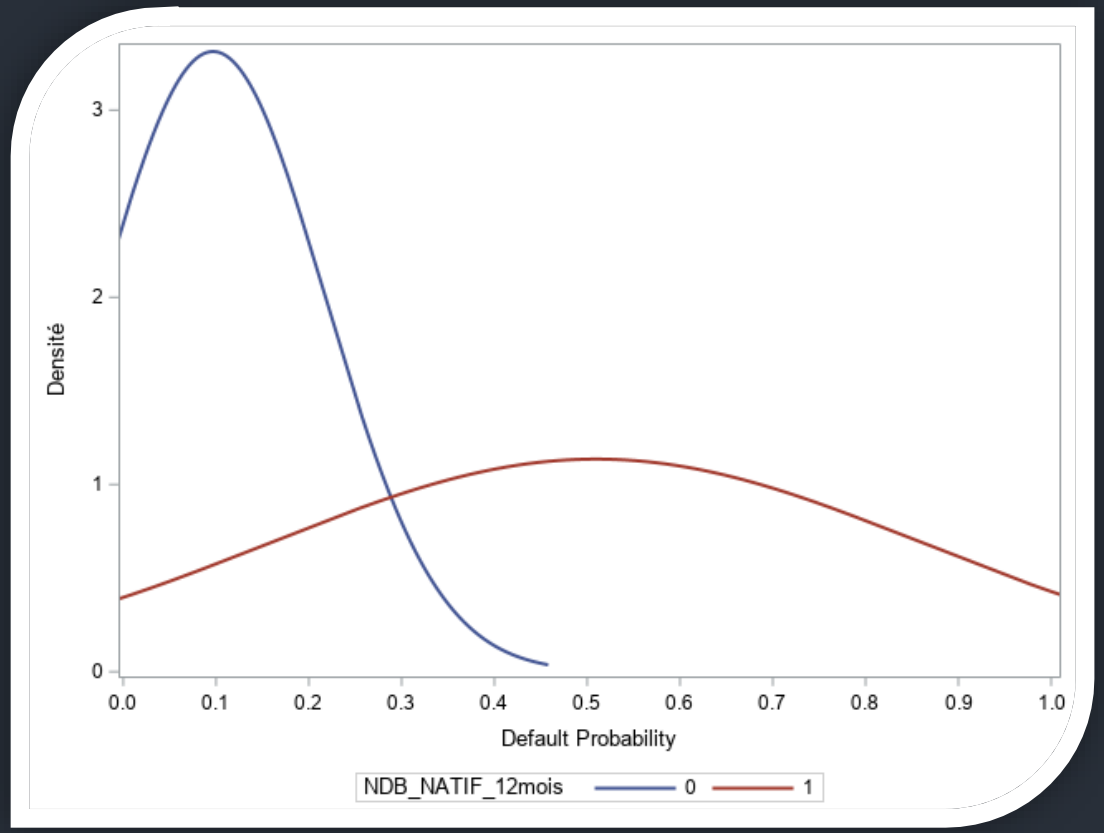
4) Courbe LIFT



Le Lift à 10% pour les défauts est de 5,5 environ : en utilisant notre modèle pour prédire les défauts en retenant les 10% des clients avec la probabilité de défaut la plus élevée, nous nous attendons à retrouver 5,5 fois plus de défauts que si l'on effectuait une requête aléatoire.

MÉTHODE III – GRADIENT BOOSTING

Classe de risque	Probabilité de défaut
Peu risquée	[0,0.12]
Assez risquée	[0.12,0.34]
Risquée	[0.34,0.665]
Très risquée	[0.665,1]



Densités conditionnelles Gradient Boosting

Le modèle semble très discriminant car nos deux courbes de densité sont bien éloignées l'une de l'autre. Ces densités sont très proches de celle du Random Forest.

Méthode III – GRADIENT BOOSTING

Avantages du Gradient Boosting :

- L’algorithme est simple à mettre en place et fournit une interprétabilité relativement simple grâce à la représentation sous forme d’arbre.
- Cette méthode limite les risques de sur-apprentissage.

Inconvénient du Gradient Boosting :

- L’algorithme est sensible aux valeurs aberrantes puisque chaque classifieur est obligé de corriger les erreurs des prédécesseurs.

XGBoost

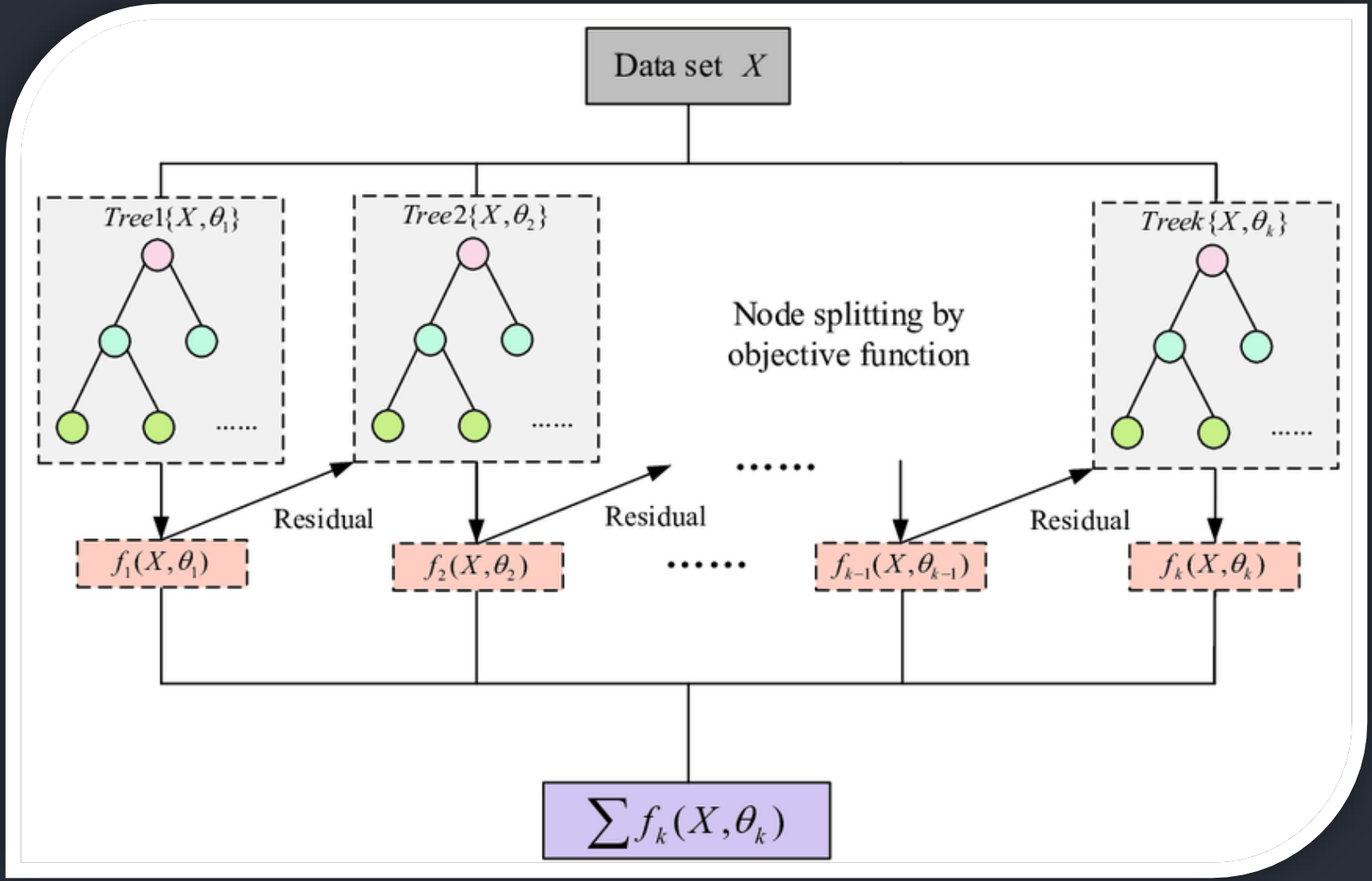
Méthode IV – XGBOOST

XGBoost signifie **eXtreme Gradient Boosting**. Comme son nom l'indique, il s'agit d'un algorithme de Gradient Boosting amélioré.

La principale différence entre le XGBoost et la méthode du Gradient Boosting réside dans le fait que le XGBoost est **informatiquement optimisé** pour rendre les différents calculs nécessaires à l'application d'un Gradient Boosting plus rapide. En effet, le XGBoost traite les données en plusieurs blocs compressés permettant de les trier beaucoup plus rapidement ainsi que **de les traiter en parallèle.**

De plus, grâce au large panel d'hyperparamètres proposé par l'algorithme du XGBoost, il est possible d'avoir un contrôle total sur l'implémentation du Gradient Boosting et de rajouter différentes régularisations dans la fonction de perte.

Méthode IV – XGBOOST



Algorithme du XGBoost

Méthode IV – XGBOOST

1) Matrice de Confusion

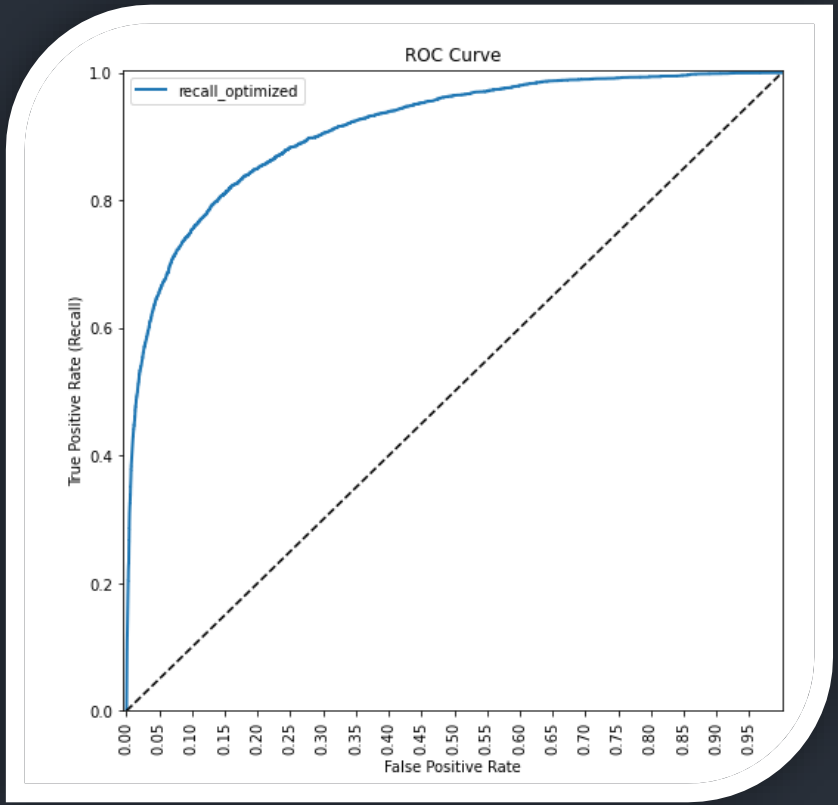
	0 prédits	1 prédits
0 observés	26434	4019
1 observés	1230	4785

Sensibilité : 86,80 %

Spécificité : 79,55 %

Méthode IV – XGBOOST

2) Courbe ROC



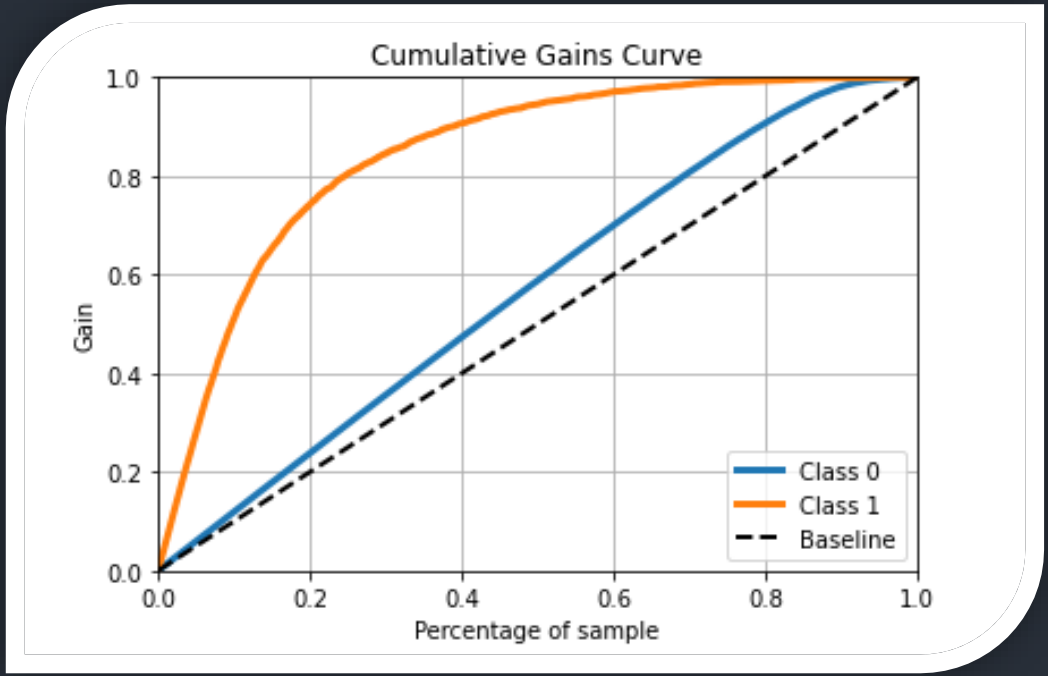
Area-Under-the-Curve : 0,91

Indice de Gini : 0,82

Si l'on choisit deux clients au hasard, un défaillant et un sain, il y a 91% de chance pour que la probabilité de défaut soit plus élevée pour l'individu réellement en défaut.

Méthode IV – XGBOOST

3) Courbe des gains cumulés

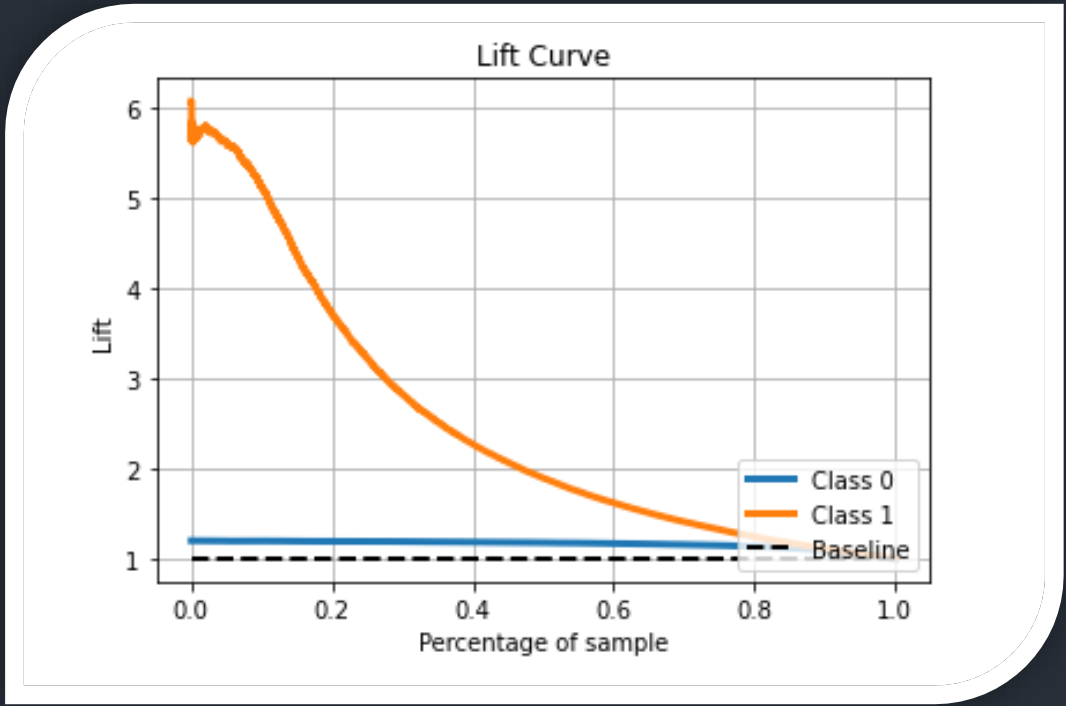


La ligne de référence pointillée représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

Ici, nous voyons que 20% des individus de la base associés aux probabilités de défaut les plus élevées contiendraient environ 75% des clients défaillants.

Méthode IV – XGBOOST

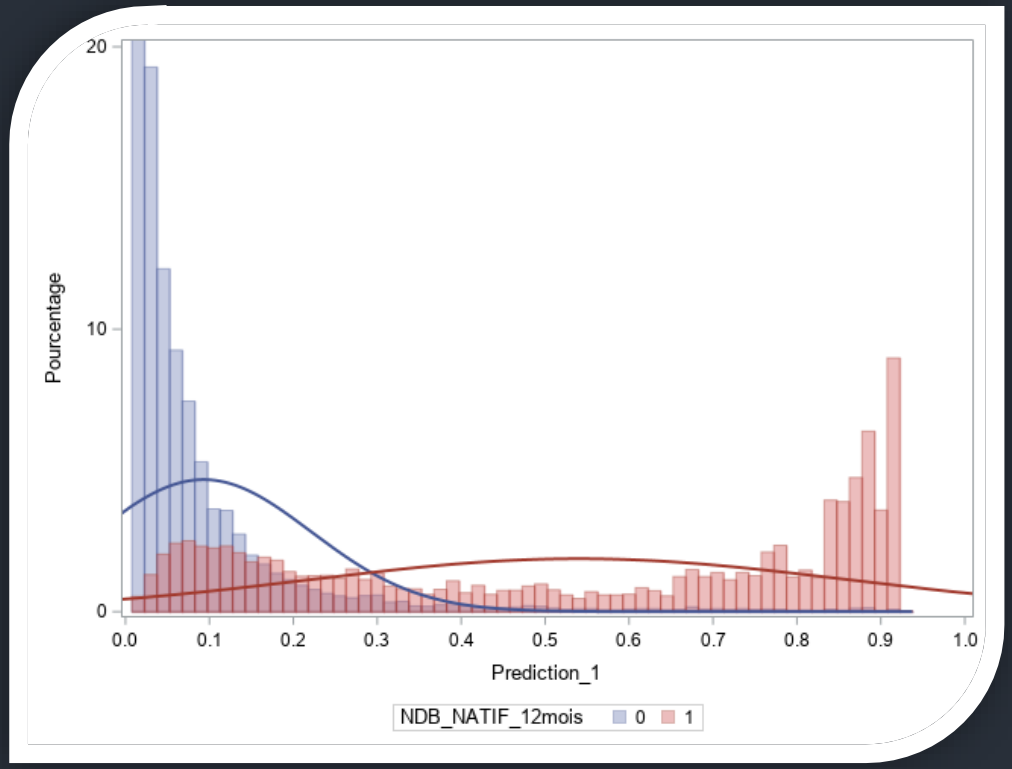
4) Courbe LIFT



Le Lift à 10% pour les défauts est de 5,5 environ : en utilisant notre modèle pour prédire les défauts en retenant les 10% des clients avec la probabilité de défaut la plus élevée, nous nous attendons à retrouver 5,5 fois plus de défauts que si l'on effectuait une requête aléatoire.

Méthode IV – XGBOOST

Classe de risque	Probabilité de défaut
Peu risquée	[0,0.11]
Assez risquée	[0.11,0.32]
Risquée	[0.32,0.643]
Très risquée	[0.643,1]



Densités conditionnelles XGBoost

Le modèle semble très discriminant car nos deux courbes de densité sont très éloignées l'une de l'autre. Notre modèle est donc très performant pour classer nos individus.

Méthode IV – XGBOOST

Avantages du XGBoost :

- L’algorithme est plus rapide, plus précis et plus efficace que l’algorithme du Gradient Boosting
- Palie au risque d’overfitting de l’algorithme du Gradient Boosting

Limites du XGBoost :

- Le XGBoost constitue une amélioration du Boosting. Dès lors, toute circonstance visant à limiter les performances du Gradient Boosting limite les performances du XGBoost.

Support Vector Machines

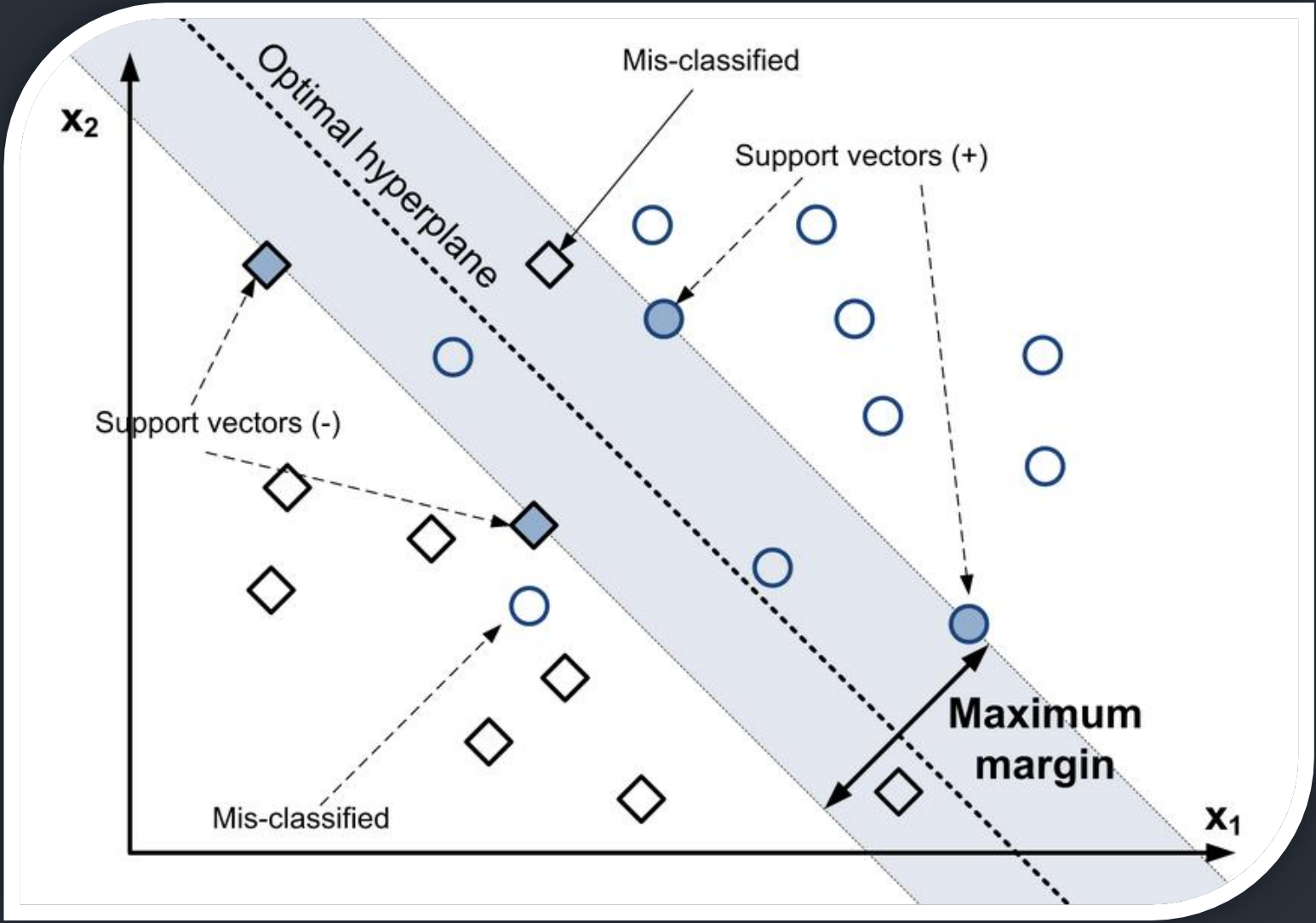
Méthode V – SVM

- Les SVMs (Support Vector Machine) sont une **famille d'algorithmes d'apprentissage automatique** qui permettent de résoudre des problèmes de classification ou de régression.

- Le principe des SVMs est simple : ils ont pour but de séparer les données en classes à l'aide d'une frontière aussi simple que possible, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale. Cette distance est appelée la « **marge** » et les points situés au plus près de cette dernière sont **les vecteurs de support**.

- La notion de frontière suppose implicitement que les données soient linéairement séparables, ce qui est rarement le cas. Pour y pallier, les SVMs reposent sur l'utilisation de **noyaux**, permettant de projeter nos données dans un espace vectoriel de plus grande dimension afin de s'assurer de la **séparabilité linéaire de nos classes**.

Méthode V – SVM



Principes du SVM

Méthode V – SVM

- Lors de l'implémentation du SVM, c'est **le Kernel linéaire** qui a été retenu pour nos données. Autrement dit, nous considérerons **un cas séparable**. Les données de notre échantillon sont donc linéairement séparables : l'espace de départ $\mathcal{F}$ est donc identique à l'espace d'arrivée du produit scalaire $\mathcal{X}$.

- Dès lors, nous nous attendons à ce que **les résultats issus de notre modèle SVM soient très similaires à ceux obtenus par une régression logistique**.

- La seule différence réside dans le **processus de l'algorithme** : en effet, au lieu de supposer un modèle probabiliste comme le modèle logistique, le SVM cherche à déterminer un hyperplan séparateur optimal particulier, dans lequel sera définit l'optimalité dans le contexte des vecteurs de support.

- C'est pourquoi pour l'application du SVM, nous n'avons retenu que les variables retenues lors de la modélisation logistique.

Méthode V – SVM

1) Matrice de Confusion

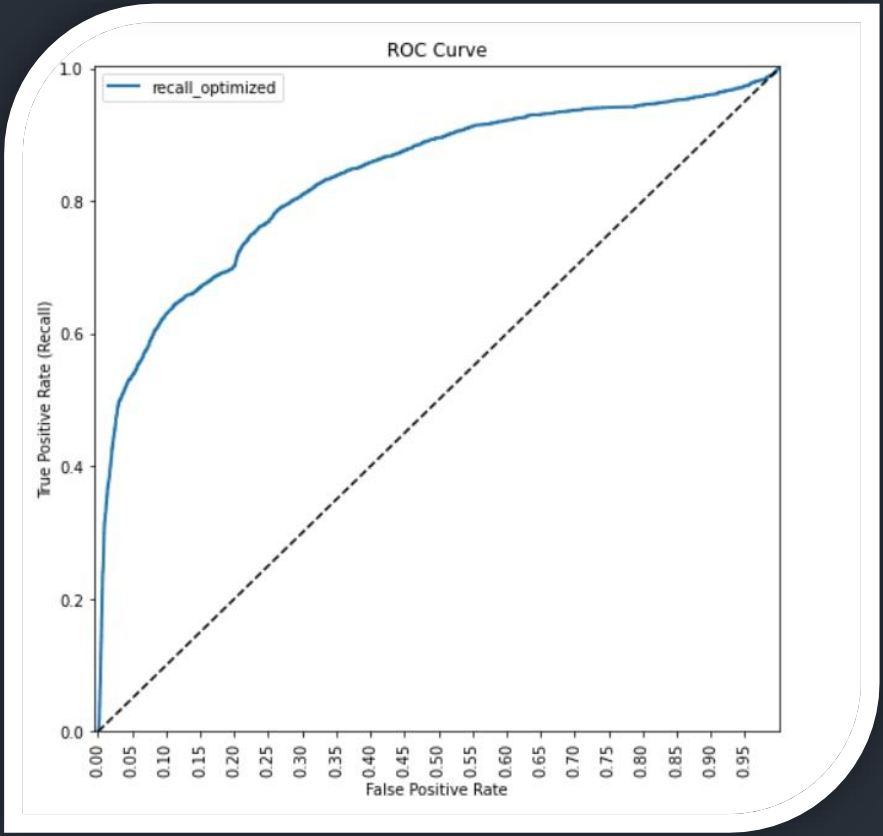
	0 prédits	1 prédits
0 observés	21391	9062
1 observés	1158	4857

Sensibilité : 80,74 %

Spécificité : 70,24 %

Méthode V – SVM

2) Courbe ROC



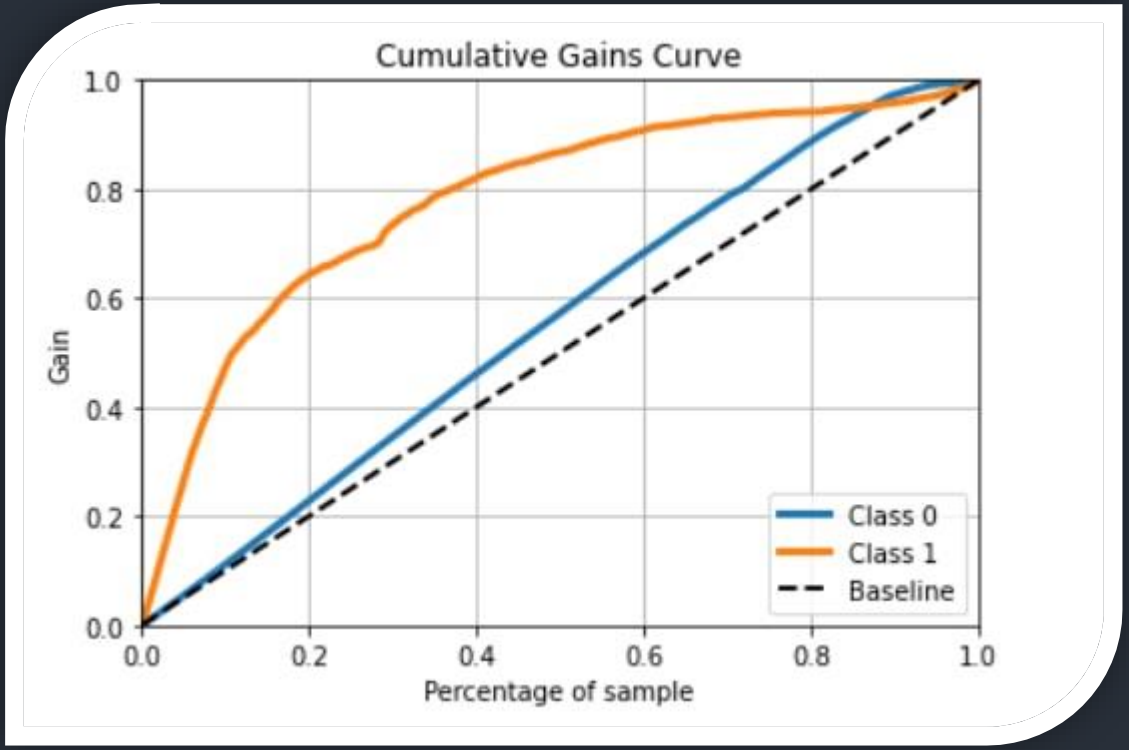
Area-Under-the-Curve : 0,84

Indice de Gini : 0,67

Si l'on choisit deux clients au hasard, un défaillant et un sain, il y a 84% de chance pour que la probabilité de défaut soit plus élevée pour l'individu réellement en défaut.

Méthode V – SVM

3) Courbe des gains cumulés

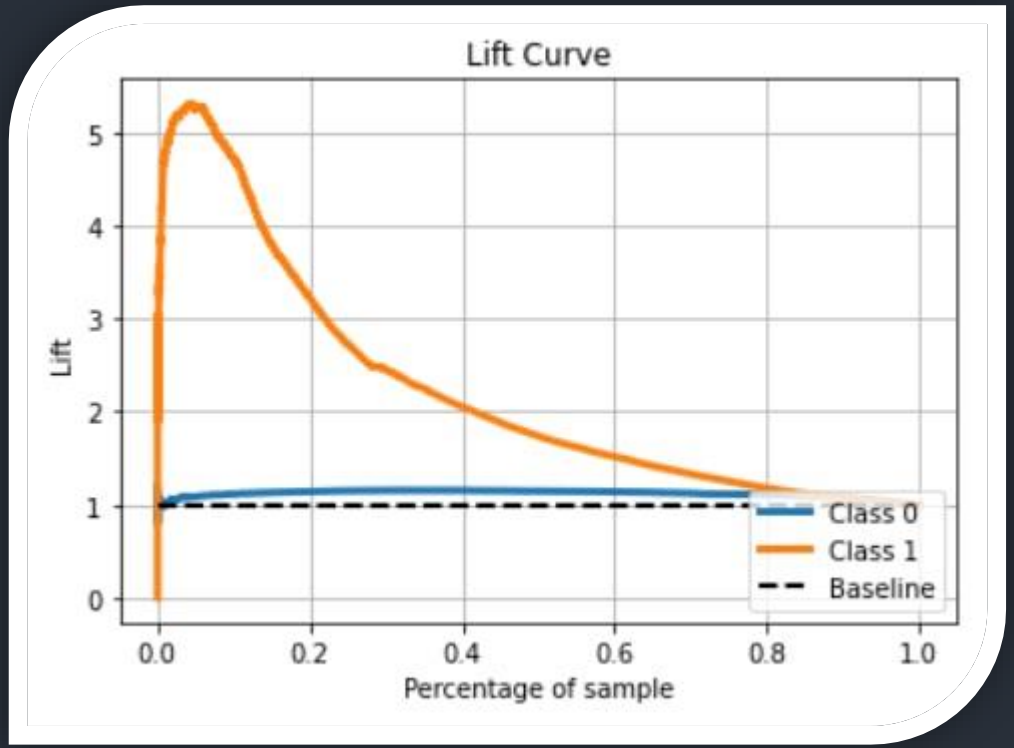


La ligne de référence pointillée représente une ligne avec une pente égale à 1, ce qui correspond à la réponse aléatoire attendue sans le modèle. Les gains supérieurs à 1 indiquent que les résultats du modèle prédictif sont meilleurs que les résultats aléatoires.

Ici, nous voyons que 20% des individus de la base associés aux probabilités de défaut les plus élevées contiendraient environ 61% des clients défaillants.

Méthode V – SVM

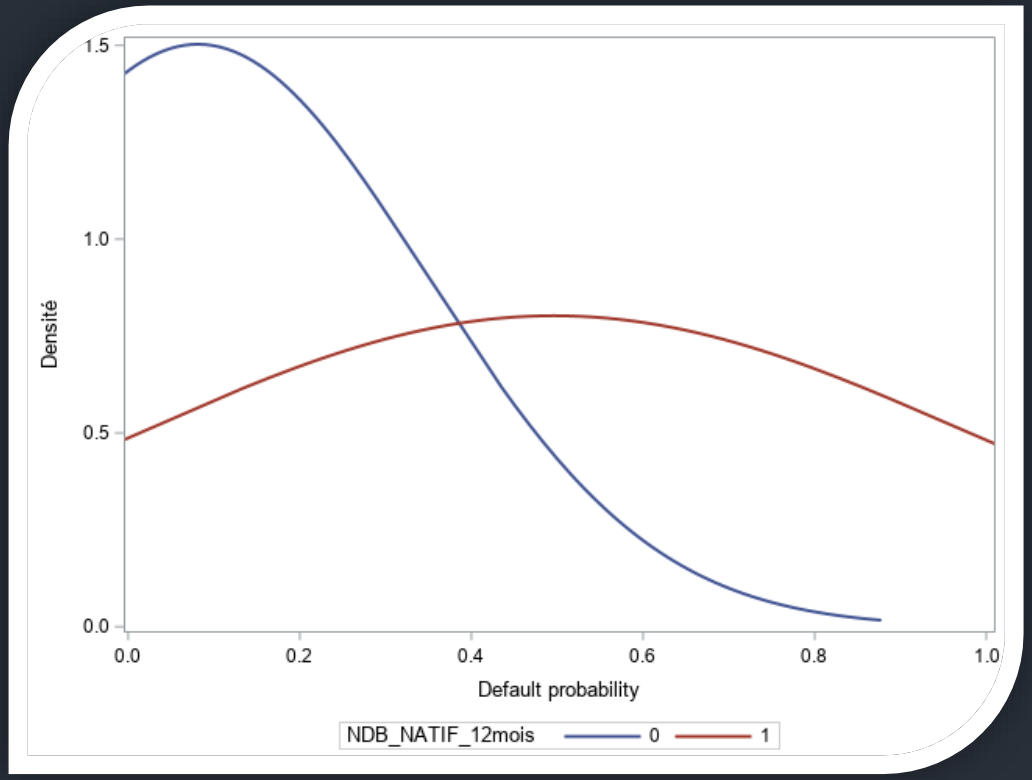
4) Courbe LIFT



Le Lift à 10% pour les défauts est de 5 environ : en utilisant notre modèle pour prédire les défauts en retenant les 10% des clients avec la probabilité de défaut la plus élevée, nous nous attendons à retrouver 5 fois plus de défauts que si l'on effectuait une requête aléatoire.

Méthode V – SVM

Classe de risque	Probabilité de défaut
Peu risquée	[0,0.0785]
Assez risquée	[0.0785,0.2983]
Risquée	[0.2983,0.7188]
Très risquée	[0.7212,1]



Densités conditionnelles SVM

Le modèle semble assez discriminant car nos deux courbes de densité sont assez éloignées l'une de l'autre.

Méthode V – SVM

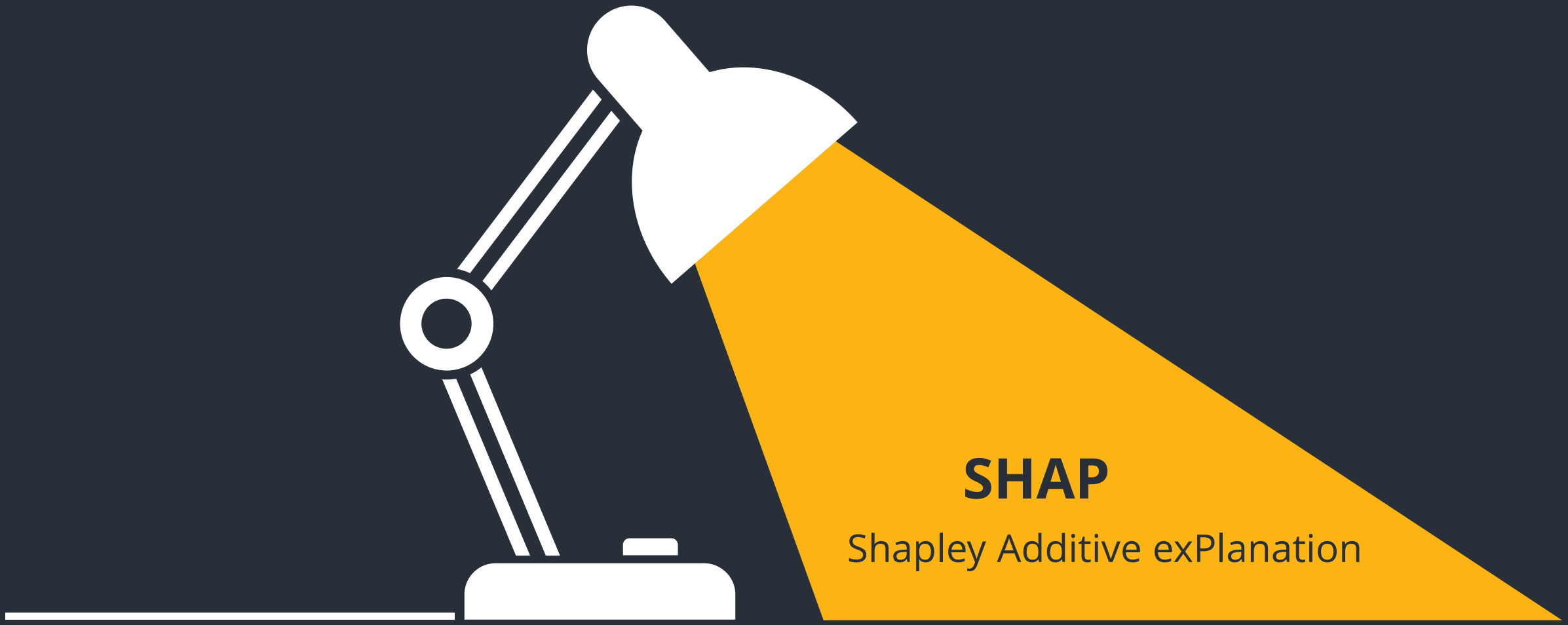
Avantages du SVM :

- Le paramètre de régularisation permet d'éviter les sur-ajustements.
- Le SVM est défini par un problème d'optimisation convexe. Il n'existe donc pas de minima locaux.

Inconvénients du SVM :

- La fonction de perte utilisée pour la régression des vecteurs de support n'a pas d'interprétation statistique évidente.
- Les modèles de noyau peuvent être assez sensibles au sur-ajustement du critère de sélection du modèle.
- Les temps de calculs de l'algorithme peuvent augmenter très fortement, en fonction de la taille de la base de données fournit

MACHINE LEARNING INTERPRÉTABLE



- Dans cette partie, nous nous focaliserons sur deux modèles, à savoir le XGBoost et le SVM.

- Bien que ces modèles surperforment notre modélisation logistique, ils constituent une « boîte noire » et sont moins interprétables que le simple modèle logistique.

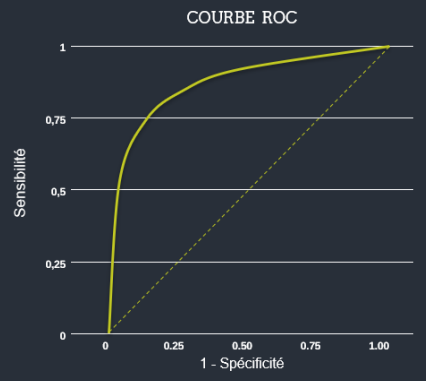
- Cette deuxième partie sera donc consacrée à l’interprétation de ces modèles. Nous utiliserons la méthode des SHAP Values, se basant sur l’importance des variables ainsi que sur la contribution de chacune d’entre elles dans notre prédiction.

PRIVATE

Comparaison des Modèles

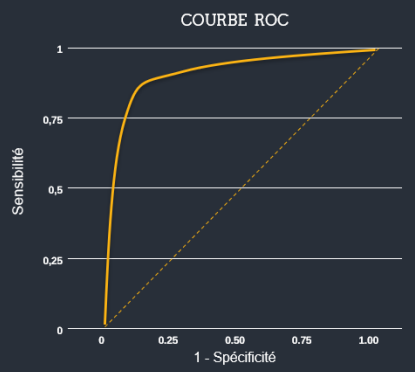
AUC
0.85
Indice de Gini
0.69

K-Near-Neighbors



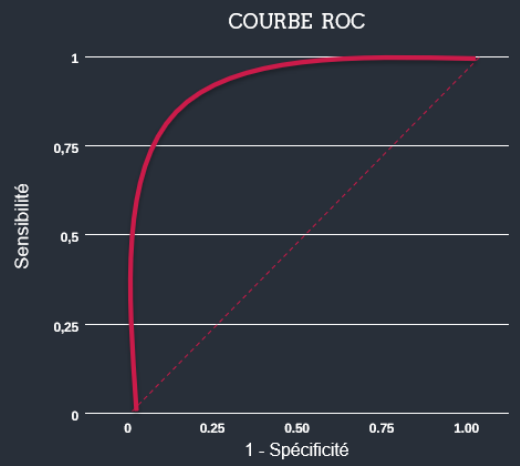
AUC
0.86
Indice de Gini
0.73

Random Forest



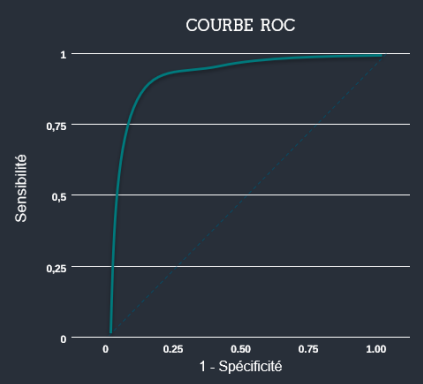
AUC
0.92
Indice de Gini
0.82

XGBoost



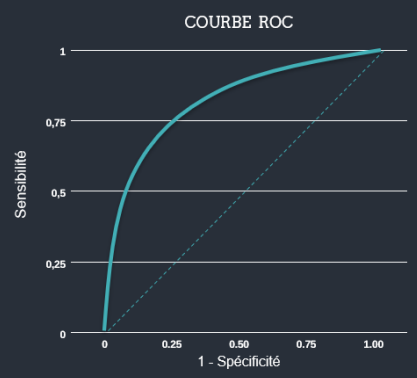
AUC
0.91
Indice de Gini
0.82

Gradient Boosting

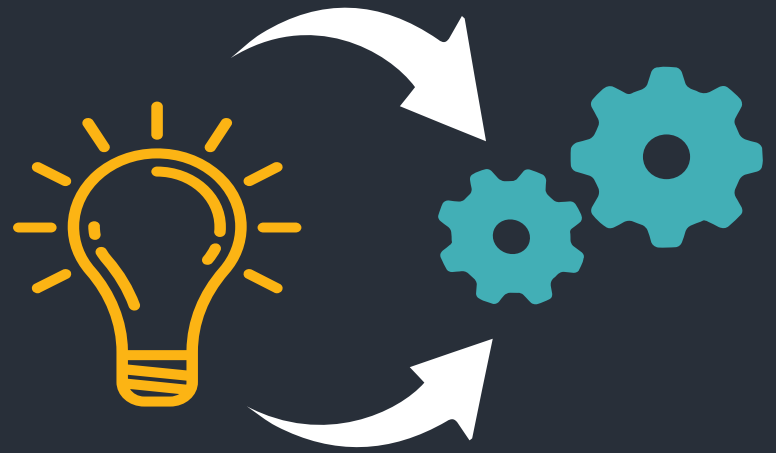


AUC
0.84
Indice de Gini
0.67

Support Vector Machine



Conclusion



- Parmi tous nos modèles, le XGBoost est celui qui s'avère être le plus performant. Malgré les outils de type SHAP, les interprétations individuelles de ce modèle peuvent différer fortement de l'interprétation globale. Par ailleurs, les interprétations issues des SHAP ne sont pas très intuitives. D'autre part, le temps de calcul peut fortement augmenter.

- Face à des modèles de Machine Learning poussés et complexes, notre modèle simple de régression logistique reste très acceptable. Cette dernière a l'avantage de permettre la construction d'une grille de score, et facilite donc très fortement les explications éventuelles quant au refus d'un prêt à un client par exemple.

- Il peut donc être intéressant de conserver le modèle de régression logistique, et d'utiliser les modèles de Machine Learning plus avancés comme des Benchmarks.

- Remarquons également que si les performances des modèles de Machine Learning et du modèle logistique sont proches, cela peut être dû au fait que nous avons ici une séparabilité linéaire (SVM linéaire). Si l'on a toutefois de fortes suspicions quant à cette linéarité, nous pouvons nous attendre à un écart de performance beaucoup plus important en faveur de nos modèles de Machine Learning, capables de capter des effets fortement non linéaires.

Sources Seconde Partie

Sources internes :

- *Cours de Machine Learning Interprétable – Mr Hurlin*
- *Cours de Support Vector Machine - Mr.Hurlin*
- *Cours de Classification – Mr. Lahiani*
- *Cours de Machine Learning - Mr.Tokpavi*

Sources externes :

- *https://datascientest.com/knn*
- *https://xgboost.readthedocs.io/en/latest/treemethod.html*
- *https://www.kaggle.com/questions-and-answers/77947*
- *https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf*
- *https://www.kaggle.com/diegovicente/using-shap-values-for-interpretability*
- *https://paulvanderlaken.com/2020/01/20/animated-machine-learning-classifiers/*

Nous adressons nos remerciements les plus sincères à Mr HURLIN ainsi qu’à Mr HUDEBINE pour son intervention.

MASTER ESA
Université d'Orléans