

PROJET 4

CONSTRUCTION D'UN MODÈLE DE SCORING: BON OU MAUVAIS PAYEUR? MODÉLISATION SUPERVISEE

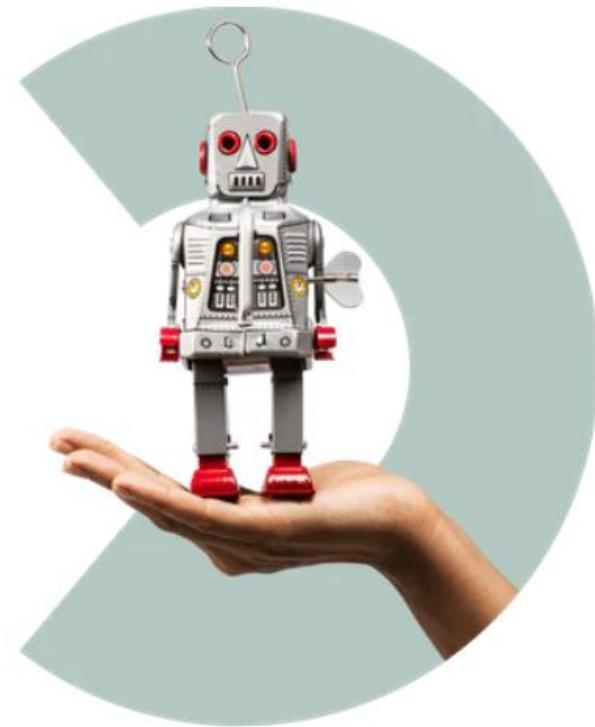
#DONNEES UNBALANCED #UNDER/OVERSAMPLING
#FEATURES ENGINEERING #FONCTION DE SCORING BESPOKE
#HYPERPARAMTRES #GRIDSEARCHCV #ROC CURVE
#REGRESSION LOGISTIQUE #SVM #BAGGING #RANDOM FOREST
#GRADIENT BOOSTING #INTERPRETABILITE
#PANDAS #NUMPY #SKLEARN #SCIPY STATS #SEABORN

Ingénieur IA

Développez et intégrez des algorithmes de Deep Learning au sein d'un produit IA

OPENCLASSROOMS

OUDDANE NABIL



SOMMAIRE

Projet 4
Soutenance

Construisez un modèle de scoring

A. INTRODUCTION

1. Contexte
2. Objectifs

B. Ressources complémentaires

1. Kaggle Kernel
2. Interprétation en machine learning

C. Projet: Peut-on prédire si le client sera capable de rembourser un crédit?

1. Contexte: Bon ou Mauvais payeur?
2. Schéma relationnel des tables: analyse des clés
3. Analyse exploratoire
4. Création de facteur .
5. MODÉLISATION .

D. ANNEXES

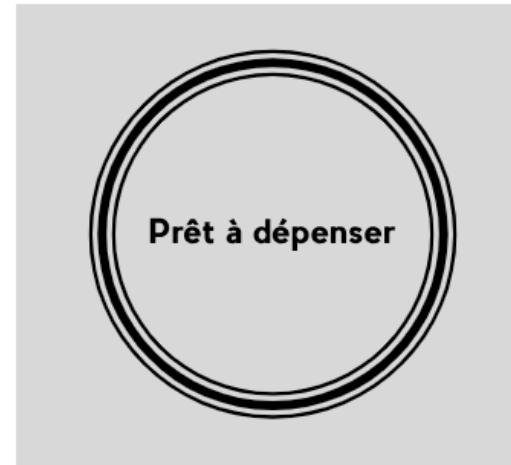
1. ANALYSE DESCRIPTIVE: CATEGORIELLES
2. ANALYSE DESCRIPTIVE: QUANTITATIVES ORDINALES
3. ANALYSE DESCRIPTIVE: QUANTITATIVES CONTINUES
4. ANALYSE DESCRIPTIVE: QUANTITATIVES CONTINUES :FACTEURS CRÉÉS DEPUIS fichier APP TRAIN GRIDSEARCHCV

A

INTRODUCTION

1. Contexte

- ENJEU global: développer un **algorithme de scoring** pour aider à décider si un prêt peut être accordé à un client
- Le modèle doit être **facilement interprétable** en affichant **une mesure de l'importance des variables** qui ont poussé le modèle à donner cette probabilité à un client
- ENJEU DU P4: Modélisation supervisée classique
- DONNEES SOURCES
 - le jeu de données :
 - https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientis_t/Projet+-+Impl%C3%A9menter+un+mod%C3%A8le+de+scoring/Projet+Mise+en+prod+-+home-credit-default-risk.zip



2. Objectifs

- SCRIPT:
 - Transformer les variables pertinentes pour un modèle supervisé classique:
 - **Construction** d'au moins trois nouvelles variables à partir des variables existantes, semblant pertinentes pour améliorer le pouvoir prédictif du modèle.
 - Entraîner un modèle supervisé classique qui répond aux attentes des métiers
 - Adapter les hyperparamètres d'un modèle d'apprentissage supervisé classique
 - Evaluer les performances d'un modèle supervisé classique



kaggle™

B

RESSOURCES COMPLÉMENTAIRES

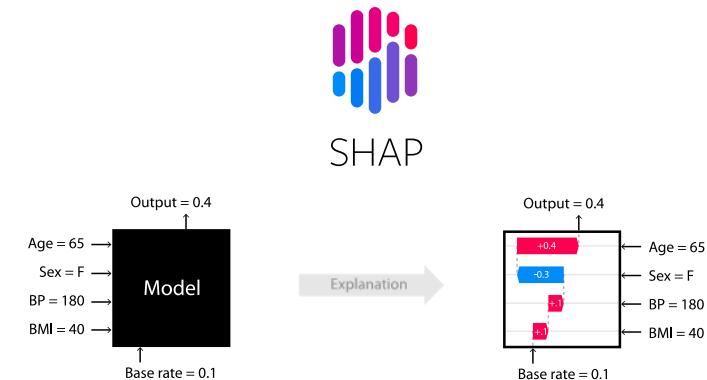
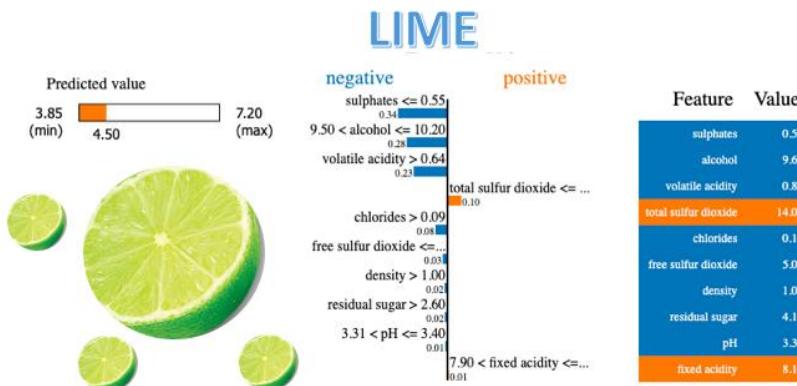
1. Kaggle Kernel

- Service cloud
 - **Kaggle** est une plateforme web, fondé en 2010, organisant des compétitions en science des données. Sur cette plateforme, les entreprises proposent des problèmes en science des données et offrent un prix aux *datalogistes* obtenant les meilleures performances
 - **Kaggle Kernel** est un plan de travail (workbench) permettant aux datascientists de partager des extraits de code (kernel/ code snippets) en R/Python. Plusieurs centaines de milliers de code sont partagés à ce jour sur différents domaines allant de l'analyse de sentiment à la détection d'objet
 - **Home Credit Default Risk:**
 - <https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction>



2. Interpretation en machine learning : blog

- Blog sur la data science
 - Towards data science
 - <https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f>
 - <https://christophm.github.io/interpretable-ml-book/>
 - <https://coderzcolumn.com/tutorials/machine-learning/how-to-use-lime-to-understand-sklearn-models-predictions>
 - <https://medium.com/@ulalaparis/reposer-les-limites-de-l-expliqueabilit%C3%A9-un-guide-avanc%C3%A9-de-shap-a33813a4bbfc>
 - <https://medium.com/analytics-vidhya/shap-part-3-tree-shap-3af9bcd7cd9b>



Graphics Principles Cheat Sheet v1.0

Communication

Effective visualizations communicate complex statistical and quantitative information facilitating insight, understanding, and decision making.

But what is an effective graph?

This cheat sheet provides general guidance and points to consider.

Planning

Why
Clearly identify the purpose of the graph, e.g. to deliver a message or for exploration?

What
Identify the quantitative evidence to support the purpose

Who
Identify the intended audience (specialists, non-specialists, both) and focus the design to support their needs

Where
Adapt the design to space or formatting constraints (e.g. clinical report, slide deck or publication)

Effectiveness Ranking

A graph is a representation of data that visually encodes numerical values into attributes such as lines, symbols and colors. The Cleveland-McGill scale can be used to select the most effective attribute(s) for your purpose.

Volume	Color hue	Depth: 3d position	Color intensity	Area	Slope or Angle	Length	Position on unaligned scale	Position on common scale
Least accurate								Most accurate
volume charts	poorly designed heat maps	multivariate density plots	heat maps	bubble charts, mosaic charts	line graphs, pie charts	stacked bar charts, waterfall chart	small multiple plots	dot plots, bar charts, parallel coordinate plots

Principles of Effective Graphic Design

Proximity – group related elements together
Alignment – elements on the same vertical or horizontal plane are perceived as having similar properties
Simplicity – cut anything superfluous, only include elements that add value, limit to 2-3 colors or fonts

White space (empty space) – use white space to minimize distraction & provide clarity

Legibility – sans serif fonts are easier to read, use color for emphasis instead of a new typeface

Color – select colors that present enough contrast to make the graph legible. Choose monochromatic color schemes to prevent clashing. Use dark colors and accent colors to emphasize important information

Visual Hierarchy – use color, font, image size, typeface, alignment & placement to create a viewing order

Focal Points – primary area of interest that immediately attracts the eye, emphasize the most important concept and make it your focal point. Use contrasting colors to draw attention

Repetition – repeating elements can be visually appealing, repeated shapes, labels, colors

Familiarity – using familiar styles, icons, navigation structure makes viewers feel confident

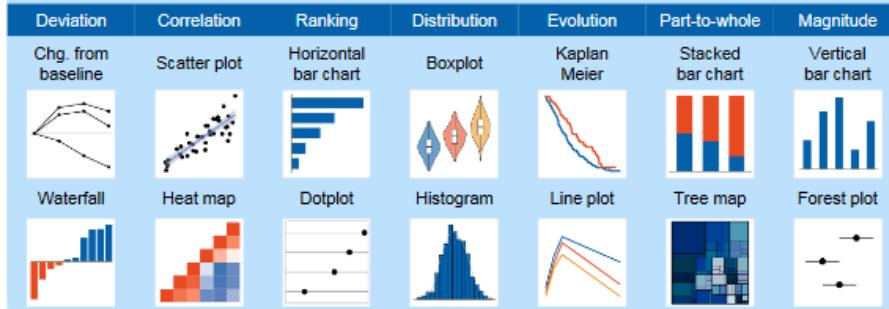
Consistency – be consistent with heading sizes, font choices, color scheme, and spacing. Use images with similar styles

Selecting the right base graph

Consider if a standard graph can be used by identifying suitable designs based on the:

(i) purpose (i.e. message to be conveyed or question to answer) and (ii) data (i.e. variables to display).

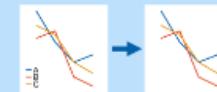
Example plots categorized by purpose



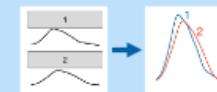
Facilitating Comparisons

Proximity improves association

Place labels next to data instead of using legends



Group together elements to be compared directly

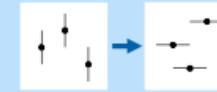


Ease visual inspection

Order values to help compare across many categories



Judgments are easier to make on a common vertical scale



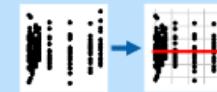
Reduce mental arithmetic

Plot the final comparison e.g. mean difference not two means

Exception: If comparator is of interest in itself



Use reference lines and other visual anchors.



Color for emphasis or distinction

Restrained use of color is highly effective in organizing a narrative and calling attention to certain elements.

Think carefully before introducing additional color. Do you really need it?

Do not use color to differentiate between categories of the same variable



Use colors or shades to represent meaningful differences such as positive/negative values, treatments or doses



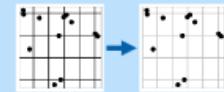
Be consistent, use the same color to mean the same thing in a series of graphs (e.g. treatment, dose)



Use a bold, saturated or contrasting color to emphasize important details.



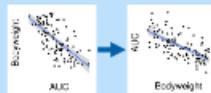
Emphasize the data by minimizing unnecessary ink, e.g. soften gridlines with a light color



Utilize existing resources for selection of appropriate palettes such as Color brewer or Munsell

Implementation Considerations

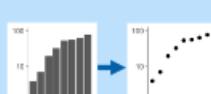
Plot cause on the x-axis and effect on the y-axis. Use this standard convention in order to avoid misinterpretation.



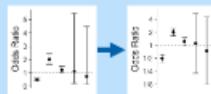
Aspect ratio can influence interpretation. Aim for a 45 degree angle of change to avoid over-interpretation of slope.



Use position for comparisons rather than length (i.e. dots instead of bars), especially for non-linear scales (e.g. log scale or % change).



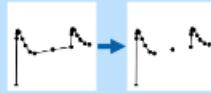
Do not plot log-normally distributed variables on a linear scale (e.g. hazard ratio, AUC, CL)



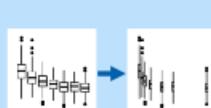
When displaying data measured on the same scale, also plot them on the same scale for easy comparison.



Connected data imply continuity. Do not connect data across a disconnected or uneven time scale.



Visits displayed close together are perceived to be closer in time. Space the visits proportional to the time between each in order to avoid confusion. Exception: baseline or pre-dose

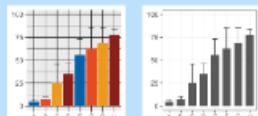


Plot data and inferences to support stories about models.

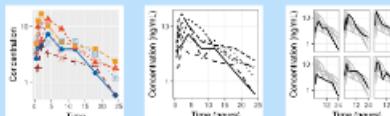


Putting it all together – Remove the clutter & emphasize the message

Creating a graph is an iterative process: produce, review and refine.



Colors, backgrounds, and borders can be removed and gridlines reduced.



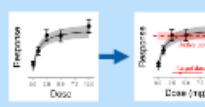
It is easier to see differences in position over a difference in length, i.e. a dot over a bar.

Using too many colors can be distracting. Use white background and try using other methods to distinguish different curves.

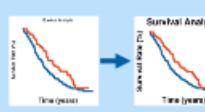
Legibility and Clarity

Effective graphs stand alone. They use titles, annotations, labels, shapes, colors, and textures to deliver important information.

Label axes with clear measurement units and provide annotations that support the message.



Use font size to create hierarchy (e.g. set titles 2pt larger than all other labels to make them more prominent).



Do not type too small or too condensed. Break long titles into two lines. Shift or adjust size of labels that overlap.



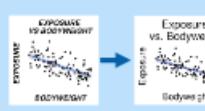
Keep the font style simple – sans serif is easier to read.



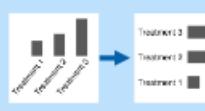
Display text with enough contrast to be visible. Favor the use of dark on light instead of light on dark whenever possible.



Bold or italics should only be used for layering or emphasis. Emphasizing everything means nothing gets emphasized.



Try not to set text at an angle, as this decreases readability. Think of alternative solutions such as transposing the graph.



Good graph checklist

Clear Communication

- Is the message of the graph as clear as possible?
- Is it easy for someone unfamiliar with the data to interpret the graph?
- Are the patterns/relationships easily identified?
- Is the graph tailored to its primary purpose and audience?
- Is the correct graph type used?

Facilitating Comparisons

- Are elements to be compared grouped together?
- Are labels placed next to data instead of in legends?
- Have categories been ordered for easy comparison?
- Can the plot be read without doing mental calculations?
- Are the estimates of interest plotted (e.g. mean differences with confidence intervals)?

Implementation Considerations

- Are multiple panels plotted on the same scale?
- Are lognormally distributed variables plotted on a log scale?
- Are common baselines used wherever possible?
- Does the orientation of the axes aid interpretation?
- Does the aspect ratio allow the reader to see variations in the data?
- Are data across a disconnected time scale kept disconnected?
- Are data spaced proportionally to the actual time interval (instead of according to visit number)?
- Are data and inferences plotted to support stories about models?
- Are number of patients by group reported if this adds context?

Legibility and Clarity

- Can all graphical elements be seen?
- Does the graph have a clear title, axis labels, annotations and data units?
- Can the font be read without eye strain or effort?
- Are sans-serif fonts used?
- Do text sizes have correct hierarchy (big to small, main text to subtext)?
- Are the elements of the graph clearly labeled (e.g. points, error bars, lines, shaded regions)?
- Are labels oriented horizontally where possible?

Resources

Books:

- E. R. Tufte, *The visual display of quantitative information*, Connecticut, Graphics Press, 2001.
- Cleveland, W.S. and McGill, Robert, *Graphical perception: theory, experimentation and application to the development of graphical methods*, JASA, Vol. 79, No. 351 – 564, 1984.
- S. Few, *Show Me The Numbers - Designing Tables and Graphs to Enlighten* (2nd Edition), Burlington, CA: Analytics Press, 2012.
- D. M. Wong, *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*, December 16, 2013.
- J. Doumont, *Trees, maps, and theorems: Effective communication for rational minds*. PRINCIPIAE.
- N. B. Robbins, *Creating More Effective Graphs*. Chart House.

- Online resources:**
- <https://www.perceptualedge.com/> (S. Few)
 - <https://www.edwardtufte.com/tufte/> (E. Tufte)
 - <http://flowingdata.com/> (N. Yau)
 - <http://www.principiae.be/> (J. Doumont)
 - <http://andrewgelman.com/> (A. Gelman)
 - <http://www.thefunctionalart.com/> (A. Cairo)
 - <http://www.nbr-graphs.com/> (N. Robbins)

Authors

Alison Margolskee, Mark Baillie, Baldrur Magnusson, Julie Jones, Marc Vandemeulebroecke



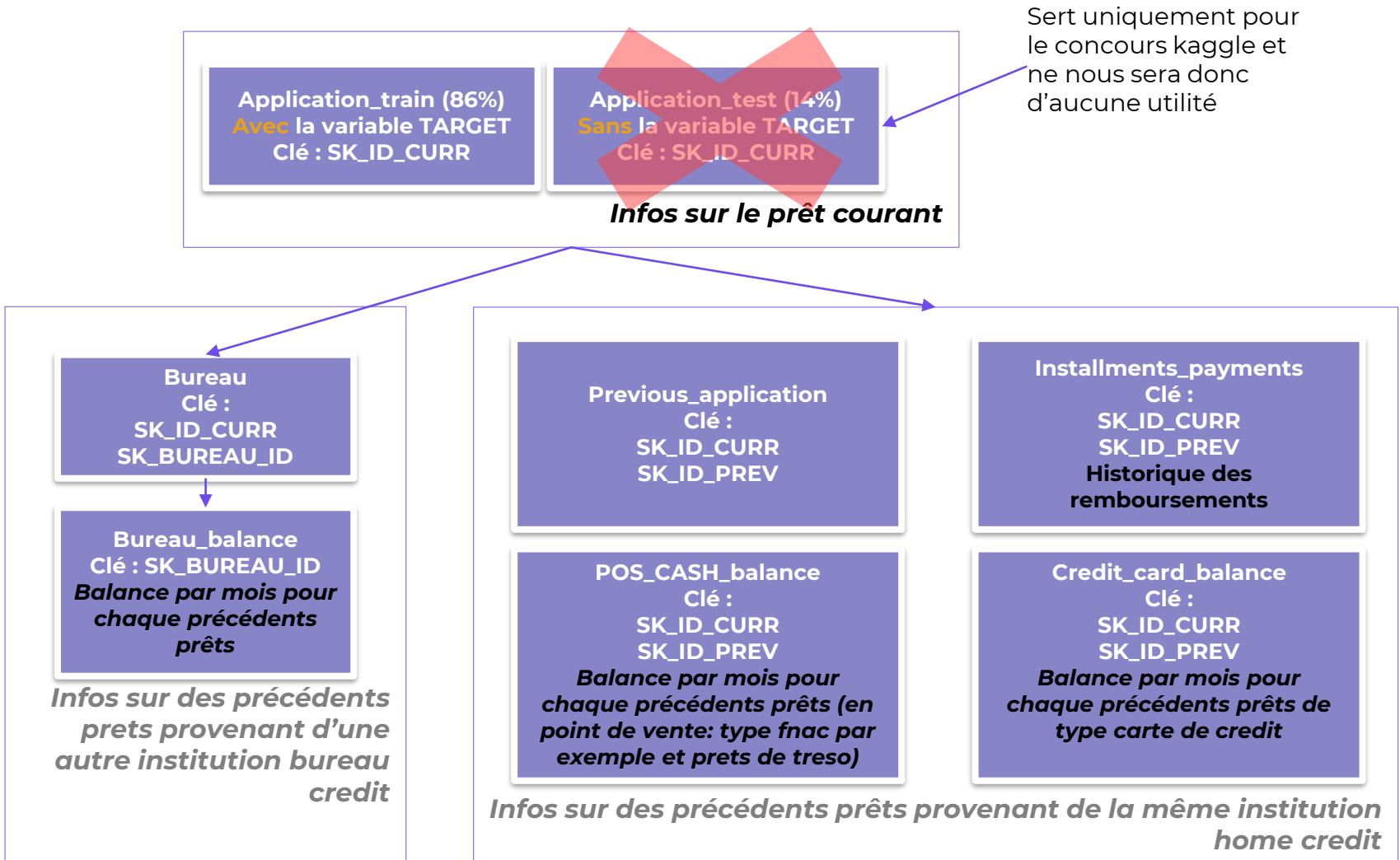
C

PROJET: *PEUT ON PRÉDIRE SI LE CLIENT SERA CAPABLE DE REMBOURSER UN CREDIT?*

1. Contexte: Bon ou Mauvais payeur?

- **Analyser les données historiques**
 - Eviter plus de défaillants avec des modèles prédictifs de machine learning
 - Imputation
 - Observation de tendances communes aux défaillants
- **Variable target**
 - 0 pour les bons payeurs capables
 - **1 pour les défaillants**
 - Données « **UNBALANCED** » : beaucoup plus de bon payeurs que de défaillants
- **Perte pour home credit**
 - Scenario 1: prédiction bon payeur qui sera défaillant
 - Plus couteux car default
 - Scenario 2: prédiction mauvais payeur qui remboursera au final
 - Moins couteux , on perd que les intérêts
- **Mesure**
 - Precision: $TP/(TP+FP)$ -> taux de vrais défaillants prédicts / taux de défaillant prédicts
 - Recall: $TP/(TP+FN)$ -> taux de vrais défaillants prédicts / taux de défaillant réels
 - F score
 - Propre Mesure bespoke

2. Schéma relationnel des tables: analyse des clés



3. Analyse exploratoire

a) Fichier principal : Application Train



- Visualisation rapide de l'état de renseignement des données**

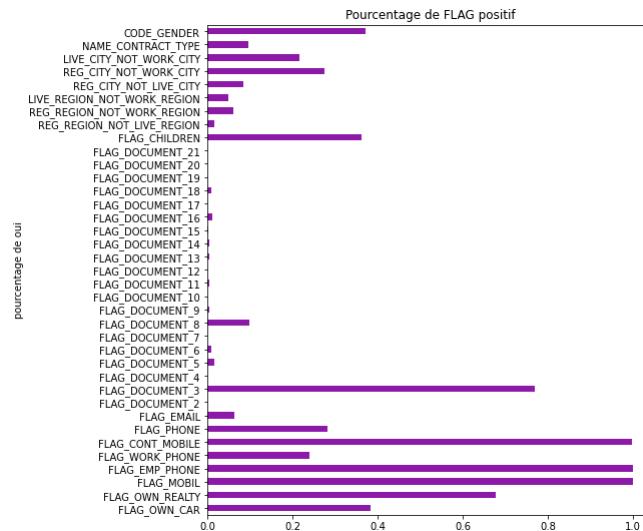
- Certains champs peu renseignés
 - les suffixes **mod /med /avg** :
 - informations normalisées sur les habitations des demandeurs
 - l'âge de la voiture:**
 - champs qui nécessite d'en posséder une pour en avoir une
 - 2/3 n'ont pas de voitures
 - 1/3 ont donné l'âge de la voiture, ça correspond
 - Profession manquante d'1/3 des demandeurs**
 - la proportion de remplissage des champs semblent identiques que le TARGET soit à 0 ou 1
- on peut donc dans un premier temps oublier un certain nombre de colonnes peu importantes**
- Et entamer un nettoyage de bon sens**

3. Analyse exploratoire

a) Fichier principal : Application Train

Nombre de modalités pour les variables catégorielles

NAME_CONTRACT_TYPE	2
CODE_GENDER	3
FLAG_OWN_CAR	2
FLAG_OWN_REALTY	2
NAME_TYPE_SUITE	7
NAME_INCOME_TYPE	7
NAME_EDUCATION_TYPE	5
NAME_FAMILY_STATUS	5
NAME_HOUSING_TYPE	6
OCCUPATION_TYPE	18
WEEKDAY_APPR_PROCESS_START	7
ORGANIZATION_TYPE	57



analyse rapide des données catégorielles

- A ce stade 12 variables catégorielles
 - 4 variables à 2 modalités peuvent subir un **LABEL ENCODING** pour devenir **QUANTITATIVES**:
 - NAME_CONTRACT_TYPE (0/1 cash/revolving)
 - CODE_GENDER (0/1 F/M) : 3^e XNA à passer en NAN
 - FLAG_OWN_CAR (0/1 N/Y)
 - FLAG_OWN_REALTY (0/1 N/Y)
 - après label encodage, il reste 8 variables catégorielles qualitatives à plus de 2 modalités

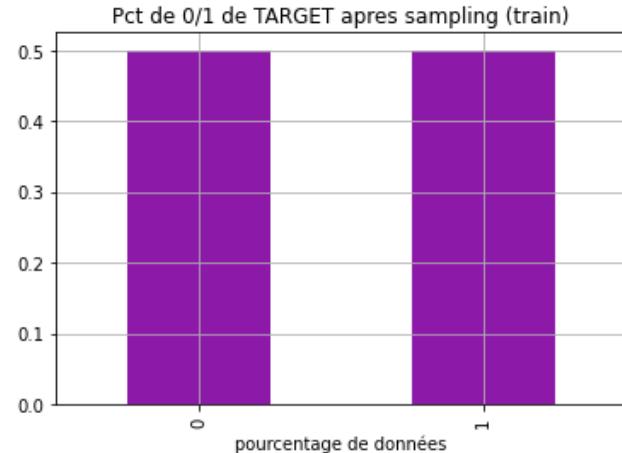
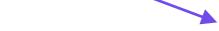
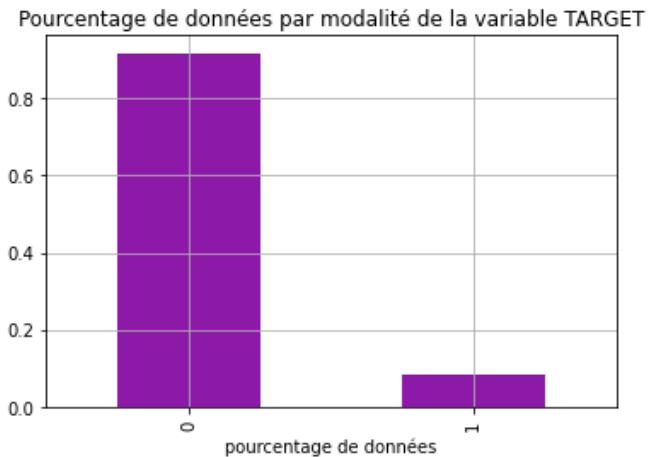
analyse rapide des données quantitatives

- 33 booléens binaires
 - TARGET et les flag ou les YES sont codés en 1
 - FLAG_MOBIL et FLAG_DOC_2 : 1 seule modalité donc INUTILES
 - A l'exception de FLAG_DOC_3 et 8, le 1 est tellement peu représenté dans les FLAG_DOC qu'il serait dangereux de les utiliser, REG_REGION_NOT_LIVE_REGION
 - le 0 est tellement peu représenté pour FLAG_CONT_MOBILE et FLAG_EMP_PHONE qu'il serait dangereux de les utiliser

3. Analyse exploratoire

a) Fichier principal : Application Train

- Analyse rapide de la variable TARGET que l'on souhaite modéliser
 - Variable **binaire**
 - 1 = Mauvais payeur / défaillance
 - 0 = Bon payeur / client
 - Variable « **UNBALANCED** »
 - Moins de 10% de flag 1 : Peu de défaillant et heureusement
- **Oversampling** de la table « Train » pour faciliter la visualisation de l'analyse exploratoire
 - Egalisation des classes
 - Etant donné le nombre de données à notre disposition on aurait également pu faire un **UNDERSAMPLING** (chose qu'on fera par la suite pour diminuer les temps de calcul)



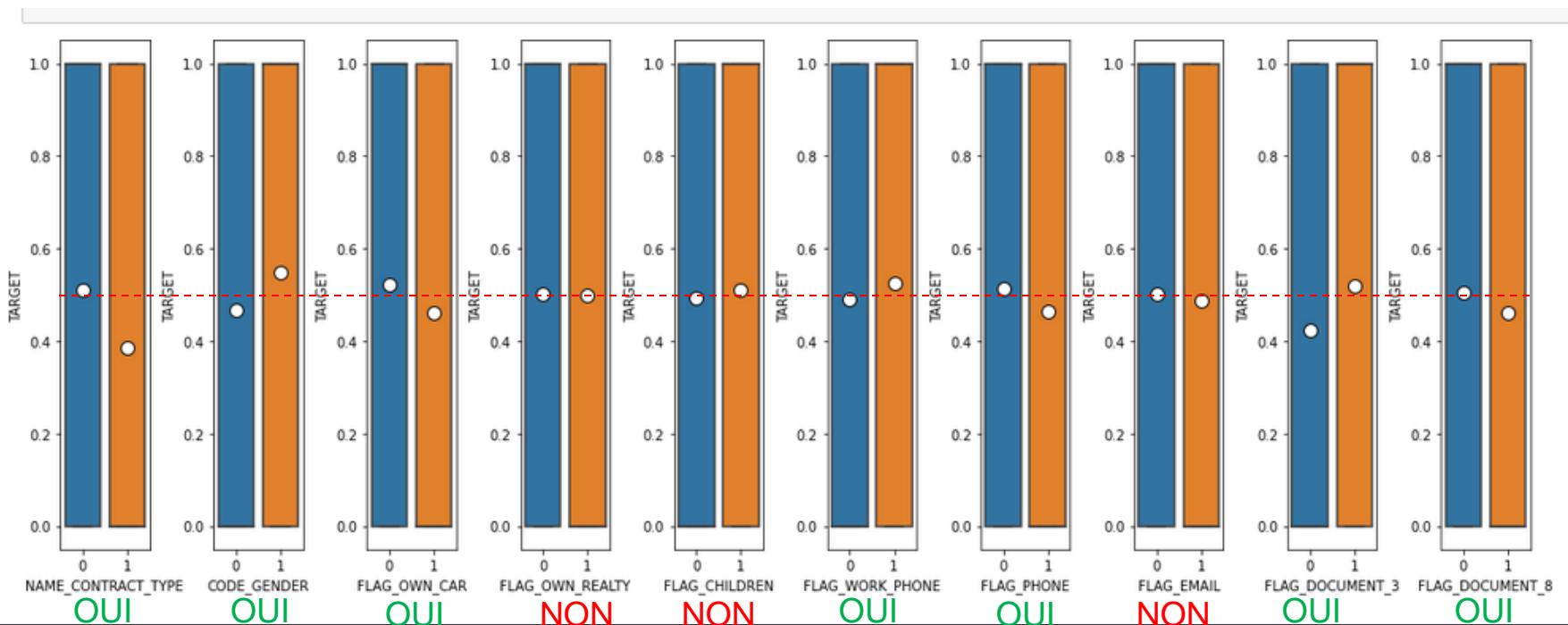
3. Analyse exploratoire

a) Fichier principal : Application Train

• ANALYSE DES RELATIONS ENTRE VARIABLES BINOMIALES ET TARGET

On peut identifier un comportement de valeur si la moyenne dans nos boxplot s'écartent de 0,5

- **contract_type**: LE REVOLVING LOANS (1) semble faire moins de défaillants
- **gender**: LES HOMMES (1) semblent être un peu plus défaillants et les femmes (0) un peu moins
- **flag car**: POSSEDER UNE VOITURE (1) rend moins défaillant
- **home phone** : donner son home phone (1) rend un peu moins défaillant mais c est léger
- **work phone** : donner son work phone (1) rend un peu plus défaillant mais c est léger
- **doc 3** :ne pas remplir le doc 3 (0) rend moins défaillant: assez clair
- **doc 8** :remplir le doc 8 (1) rend moins défaillant: assez clair
- *Les enfants , la propriété, et l'email n'apportent rien*



3. Analyse exploratoire

a) Fichier principal : Application Train

- **ANALYSE DES RELATIONS ENTRE VARIABLES CATEGORIELLES ET TARGET**
 - NAME_TYPE_SUITE / les accompagnants : **NON**
 - NAME_INCOME_TYPE / type de revenus : **OUI**
 - 3 modalités importantes
 - pas de prêts aux retraités, étudiants et femme enceintes : mais impact sur le reste,
 - les fonctionnaires sont moins défaillants
 - NAME_EDUCATION_TYPE / niveau d'étude: **OUI**
 - 3 modalités importantes
 - le niveau d'étude a clairement un impact sur le niveau de défaillance
 - NAME_FAMILY_STATUS / statut familial : **OUI** (léger)
 - NAME_HOUSING_TYPE / type d'habitation : **OUI** (léger)
 - 3 à 4 modalités importantes
 - WEEKDAY_APPR_PROCESS_START: **NON**
 - OCCUPATION_TYPE / emploi : **OUI**
 - 10 à 11 modalités importantes + imputation
 - ORGANIZATION_TYPE: **OUI**
 - Mais beaucoup de modalités , avec des pincettes

3. Analyse exploratoire

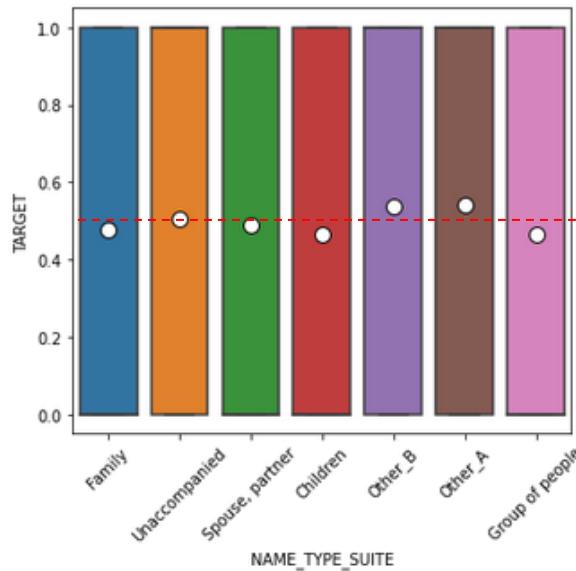
a) Fichier principal : Application Train

Unaccompanied	325741
Family	47794
Spouse, partner	14688
Children	3366
Other_B	2386
Other_A	1222
Group of people	311
Name: NAME_TYPE_SUITE, dtype: int64	

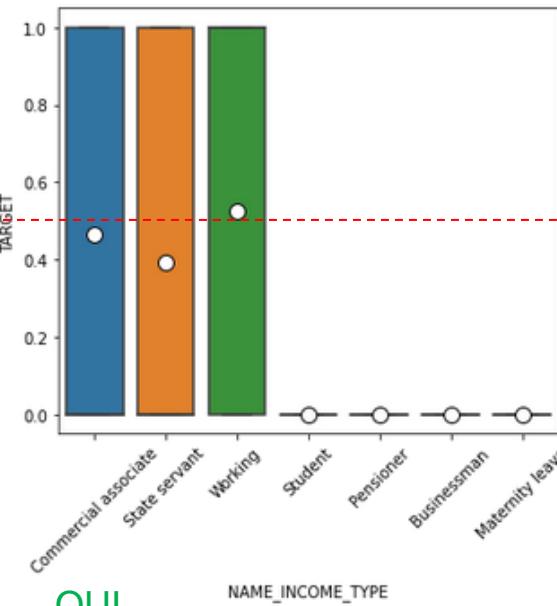
Working	259795
Commercial associate	105601
State servant	30086
Student	14
Pensioner	7
Businessman	4
Maternity leave	1
Name: NAME_INCOME_TYPE, dtype: int64	

LES ENCADREMENTS SIGNALENT DES MODALITES SUFFISAMMENT REPRESENTEES

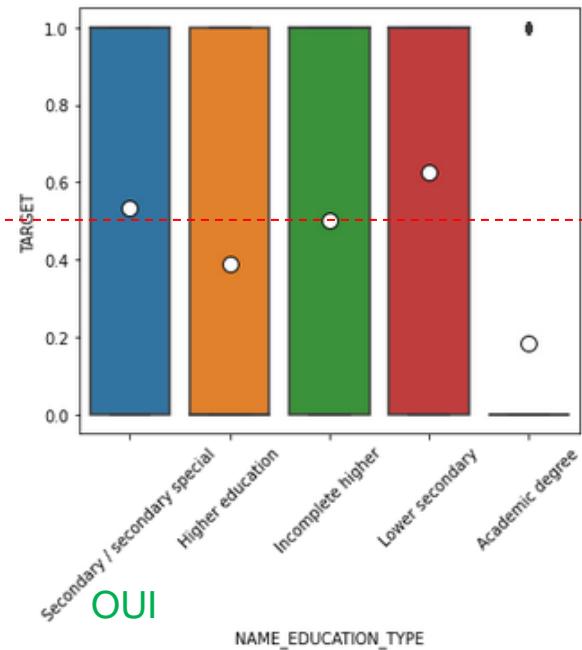
Secondary / secondary special	286188
Higher education	90031
Incomplete higher	14772
Lower secondary	4371
Academic degree	146
Name: NAME_EDUCATION_TYPE, dtype: int64	



NON



OUI



OUI

3) Analyse exploratoire

a) Fichier principal : Application Train

• ANALYSE DES RELATIONS ENTRE VARIABLES ORDINALES DISCRETES ET TARGET

- REGION_RATING_CLIENT/ note de la region d'habitation: **OUI**
REGION_RATING_CLIENT_W_CITY / avec la ville: **OUI**
 - Les 2 ratings sont très corrélés : il faut les aggrerer
- DEF_30_CNT_SOCIAL_CIRCLE / défaut dans le cercle social: **OUI**
- DEF_60_CNT_SOCIAL_CIRCLE / défaut dans le cercle social: **OUI**
 - Ces 2 variables de contexte social peuvent transformés en binaires
- OBS_30_CNT_SOCIAL_CIRCLE: **NON**
- OBS_60_CNT_SOCIAL_CIRCLE: **NON**

- AMT_REQ_CREDIT_BUREAU_YEAR : **NON**
- AMT_REQ_CREDIT_BUREAU_QRT : **NON**
- AMT_REQ_CREDIT_BUREAU_MON : **NON**
- AMT_REQ_CREDIT_BUREAU_WEEK : **NON**
- AMT_REQ_CREDIT_BUREAU_DAY : **NON**
 - Presqu'une seule modalité représentée
- AMT_REQ_CREDIT_BUREAU_HOUR : **NON**
 - Presqu'une seule modalité représentée
- Ces variables représentant le nombre de demande de renseignements sur le client avant demande de credit ne semble pas apporter d'information

3. Analyse exploratoire

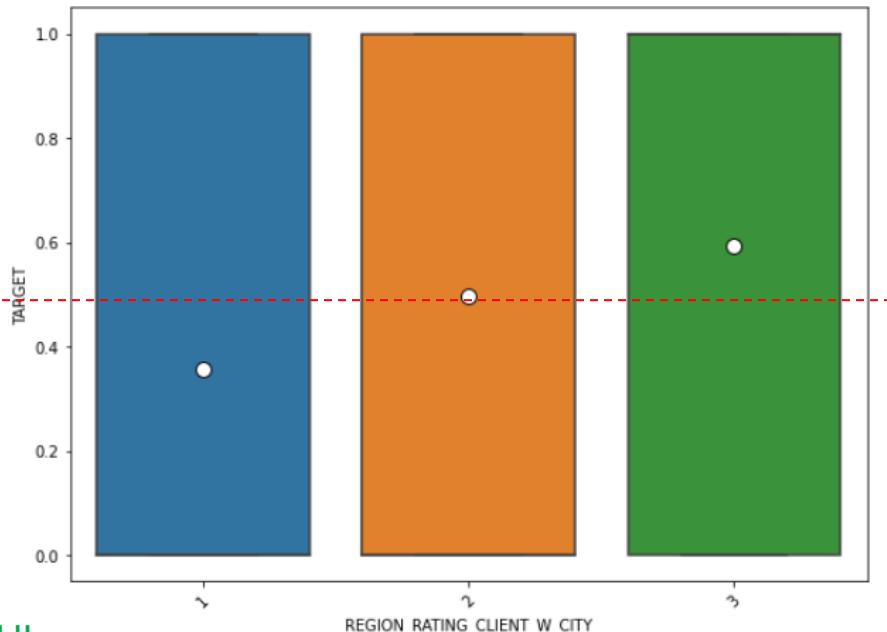
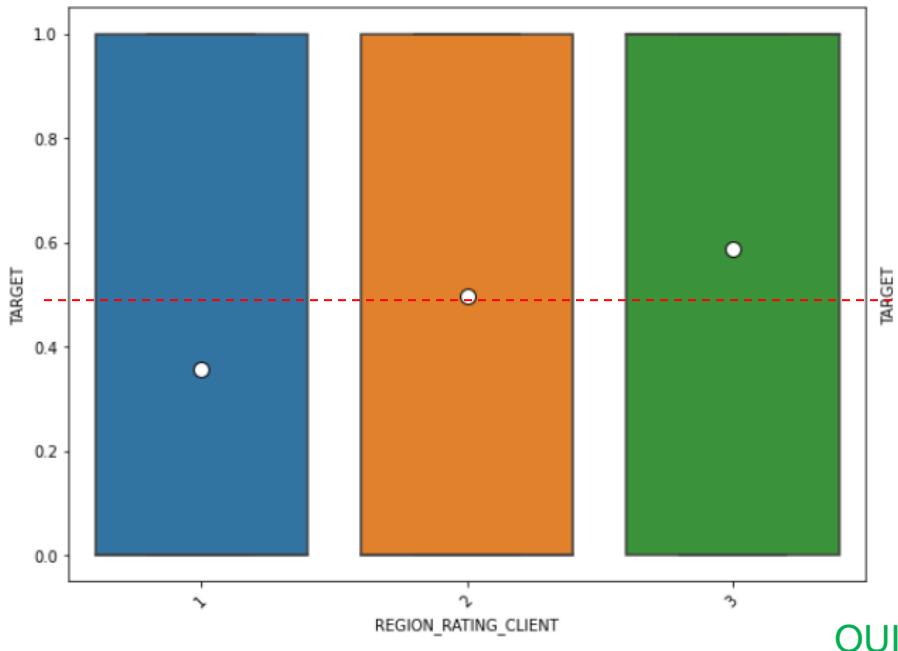
a) Fichier principal : Application Train

- Les ratings clients apportent de la valeur

```
2    290290
3    70866
1    34352
Name: REGION_RATING_CLIENT, dtype: int64
2    293755
3    65187
1    36566
Name: REGION_RATING_CLIENT_W_CITY, dtype: int64
```

Corrélation de rang de Spearman:
Correlation des ratings avec la target

	TARGET	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
TARGET	1.000000	0.109560	0.114296
REGION_RATING_CLIENT	0.109560	1.000000	0.952755
REGION_RATING_CLIENT_W_CITY	0.114296	0.952755	1.000000



3. Analyse exploratoire

a) Fichier principal : Application Train

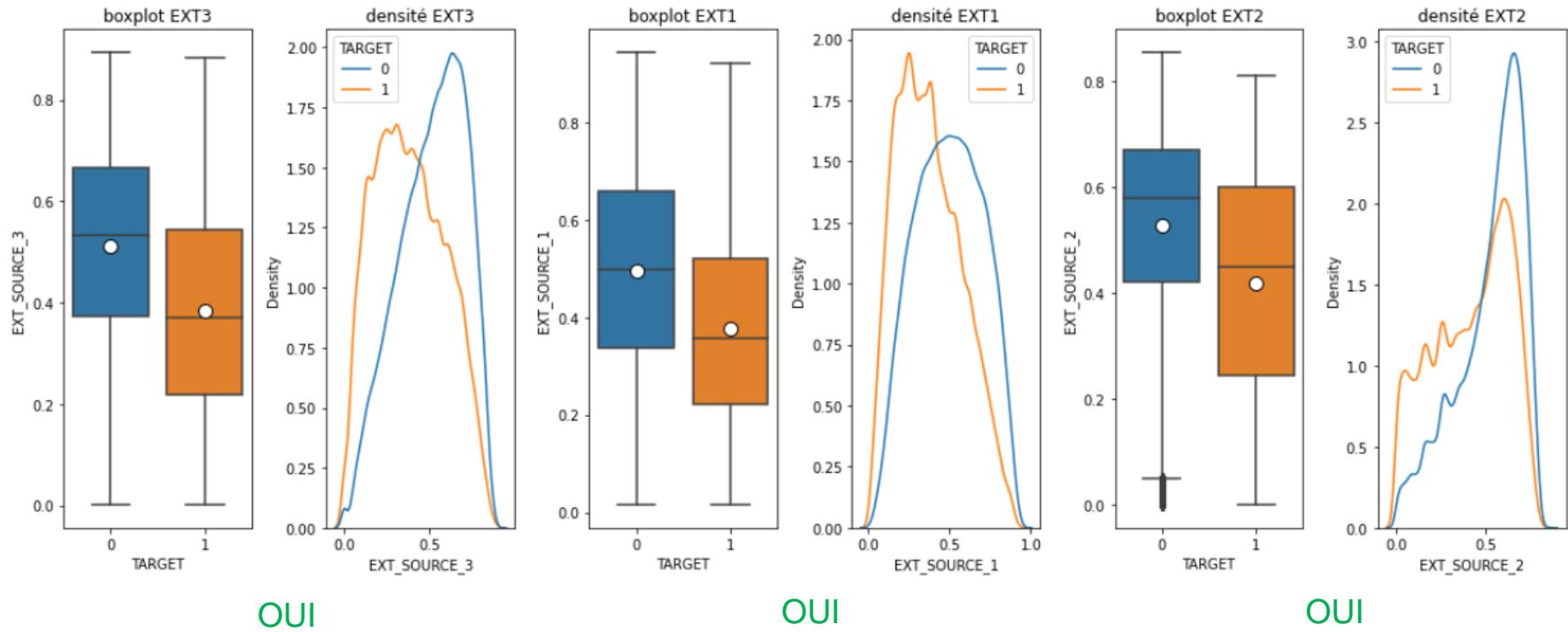
- **ANALYSE DES RELATIONS ENTRE VARIABLES QUANTITATIVES CONTINUES ET TARGET**

- EXT_SOURCE_1 / score externe **OUI**
- EXT_SOURCE_2 / score externe **OUI**
- EXT_SOURCE_3 / score externe **OUI**
 - Ces scores externes apportent de la valeur
 - Ils sont inversement proportionnel à la défaillance
 - Ils peuvent être aggregés après imputation
- DAYS_EMPLOYED / Duree d'emploi **OUI**
- DAYS_BIRTH / age **OUI**
- DAYS_LAST_PHONE_CHANGE / date de changement de telephone **PEUT ETRE**
- DAYS_ID_PUBLISH / date de changement d'information d'id **PEUT ETRE**
- AMT_INCOME_TOTAL / revenus **NON**
- AMT_CREDIT / montant du credit **NON**
- AMT_ANNUITY / annuité de remboursement **NON**
- AMT_GOODS_PRICE / prix du bien **NON**

3. Analyse exploratoire

a) Fichier principal : Application Train

- Les scores externes apportent de la valeur



3. Analyse exploratoire

a) Fichier principal : Application Train : Résumé

N°	NOM	TYPE de VARIABLES	RELATION	TRANSFORMATION A PREVOIR
1	NAME_CONTRACT_TYPE	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
2	CODE_GENDER	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
3	FLAG_OWN_CAR	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
	FLAG_OWN_REALTY	FLAG QUANTITATIVE BINAIRE	NON	
	FLAG_CHILDREN	FLAG QUANTITATIVE BINAIRE	NON	
5	FLAG_WORK_PHONE	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
6	FLAG_PHONE	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
	FLAG_EMAIL	FLAG QUANTITATIVE BINAIRE	NON	
7	FLAG_DOCUMENT_3	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
8	FLAG_DOCUMENT8	FLAG QUANTITATIVE BINAIRE	OUI	DEJA BINAIRE A CE STADE
	NAME_TYPE_SUITE	CATEGORIES	NON	
9	NAME_INCOME_TYPE	CATEGORIES	OUI	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
10	NAME_EDUCATION_TYPE	CATEGORIES	OUI	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
11	NAME_FAMILY_STATUS	CATEGORIES	OUI	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
12	NAME_HOUSING_TYPE	CATEGORIES	OUI	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
	WEEKDAY_APPR_PROCESS_START	CATEGORIES	NON	
13	OCCUPATION_TYPE	CATEGORIES	OUI	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
14	ORGANIZATION_TYPE	CATEGORIES	PEUT ETRE	LABEL OU ONE HOT ENCODER SUPPRIMER LES MODALITES LE SMOINS FREQUENTES
15	REGION_RATING_CLIENT	QUANTITATIVE DISCRETE ORDINALE	OUI	FUSION DES RATINGS
16	REGION_RATING_CLIENT_W_CITY	QUANTITATIVE DISCRETE ORDINALE	OUI	FUSION DES RATINGS
17	DEF_30_CNT_SOCIAL_CIRCLE	QUANTITATIVE DISCRETE ORDINALE	OUI	BINARISATION
18	DEF_60_CNT_SOCIAL_CIRCLE	QUANTITATIVE DISCRETE ORDINALE	OUI	BINARISATION
	OBS_30_CNT_SOCIAL_CIRCLE	QUANTITATIVE DISCRETE ORDINALE	NON	
	OBS_60_CNT_SOCIAL_CIRCLE	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_YEAR	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_QRT	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_MON	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_WEEK	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_DAY	QUANTITATIVE DISCRETE ORDINALE	NON	
	AMT_REQ_CREDIT_BUREAU_HOUR	QUANTITATIVE DISCRETE ORDINALE	NON	
19	EXT_SOURCE_1	QUANTITATIVE CONTINUE	OUI	FUSION DES EXT & imputation
20	EXT_SOURCE_2	QUANTITATIVE CONTINUE	OUI	FUSION DES EXT & imputation
21	EXT_SOURCE_3	QUANTITATIVE CONTINUE	OUI	FUSION DES EXT & imputation
22	DAYS_EMPLOYED	QUANTITATIVE CONTINUE	OUI	
23	DAYS_BIRTH	QUANTITATIVE CONTINUE	OUI	
24	DAYS_LAST_PHONE_CHANGE	QUANTITATIVE CONTINUE	PEUT ETRE	
25	DAYS_ID_PUBLISH	QUANTITATIVE CONTINUE	PEUT ETRE	
	AMT_INCOME_TOTAL	QUANTITATIVE CONTINUE	NON	
	AMT_CREDIT	QUANTITATIVE CONTINUE	NON	
	AMT_ANNUITY	QUANTITATIVE CONTINUE	NON	
	AMT_GOODS_PRICE	QUANTITATIVE CONTINUE	NON	

4. Creation de facteur

a) Fichier principal : Application Train

- **Création de nouveaux facteurs à partir du fichier principal:**

Solvabilité

- $['SOLV1']=['AMT_INCOME_TOTAL']/['AMT_ANNUITY']$
- $['SOLV2']=['AMT_INCOME_TOTAL']-['AMT_ANNUITY']$
- $['SOLV3']=(['AMT_INCOME_TOTAL']-['AMT_ANNUITY'])/['AMT_INCOME_TOTAL']$

Ages

- $['WORKAGE1']=['Y_BIRTH']-['Y_EMPLOYED']$
- $['WORKAGE2']=['Y_EMPLOYED']/['Y_BIRTH']$

Apports

- $['APPORT1']=['AMT_GOODS_PRICE']-['AMT_CREDIT']$
- $['APPORT2']=(['AMT_GOODS_PRICE']-['AMT_CREDIT'])/['AMT_CREDIT']$

Durée et durée relative aux âges

- $['DURA1']=['AMT_CREDIT']/['AMT_ANNUITY']$
- $['DURA2']=(['AMT_CREDIT']/['AMT_ANNUITY'])/['Y_BIRTH']$
- $['DURA3']=(['AMT_CREDIT']/['AMT_ANNUITY'])/['Y_BIRTH']$
- $['DURA4']=(['AMT_CREDIT']/['AMT_ANNUITY'])/(['Y_EMPLOYED']+5)$
- $['DURA5']=(['AMT_CREDIT']/['AMT_ANNUITY'])/(['Y_EMPLOYED']+5)$

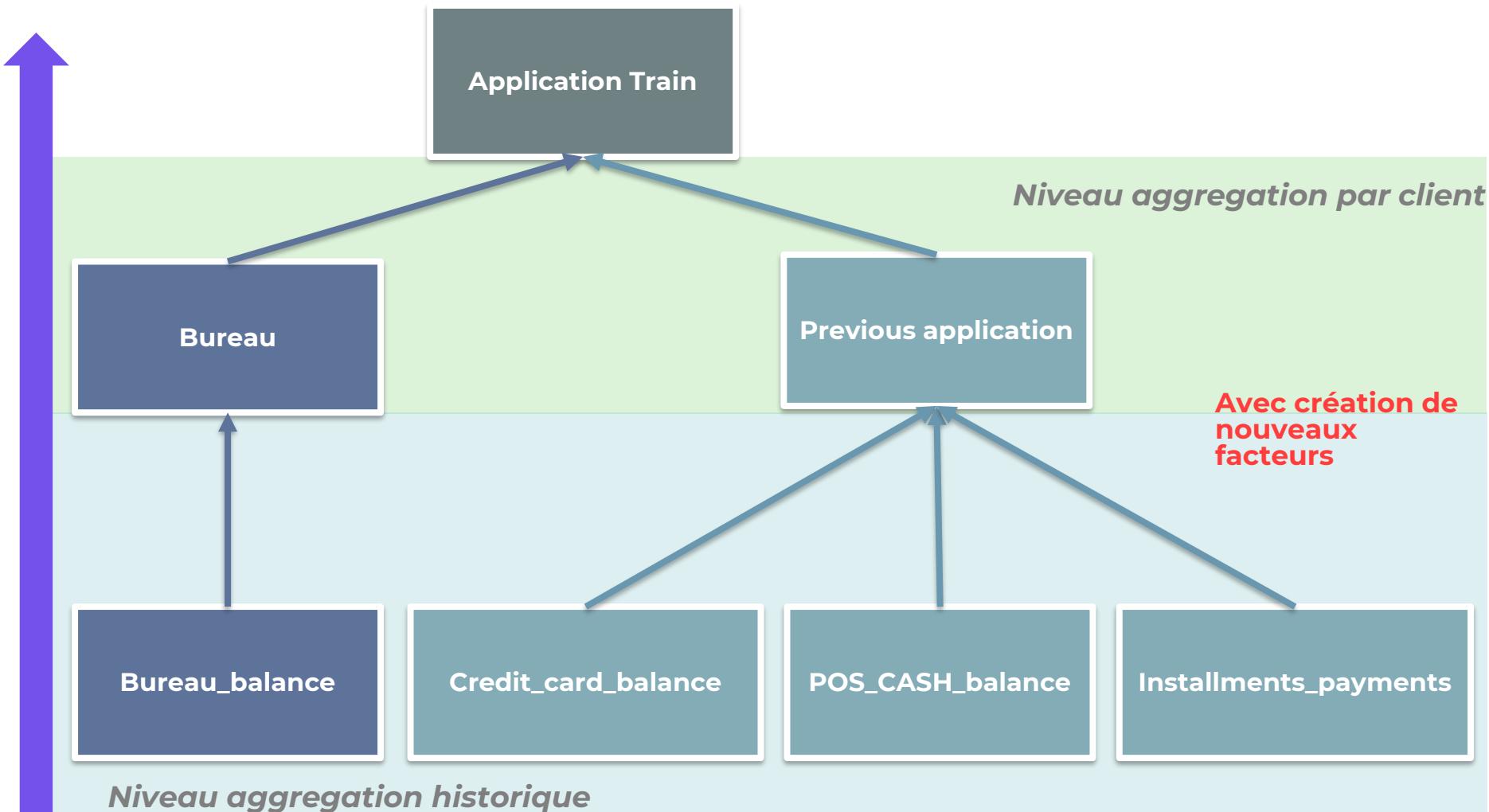
Montants relatifs aux âges

- $['SPE1']=['AMT_CREDIT']/['Y_BIRTH']$
- $['SPE2']=['AMT_ANNUITY']/['Y_BIRTH']$
- $['SPE3']=['AMT_INCOME_TOTAL']/['Y_BIRTH']$
- $['SPE4']=['AMT_CREDIT']/(['Y_EMPLOYED']+1)$
- $['SPE5']=['AMT_ANNUITY']/(['Y_EMPLOYED']+1)$
- $['SPE6']=['AMT_INCOME_TOTAL']/(['Y_EMPLOYED']+1)$

PEU DE RESULTATS EXTRAORDINAIRES AU REGARD DES GRAPHIQUES SUIVANTS (annexes) PEUT ETRE UN PEU D'ESPOIR AUTOUR DES AGES ET DES DUREES

4. Creation de facteur

b) Fichiers secondaires : AGGREGATION de variables



4. Creation de facteur

b) Fichiers secondaires

- Aggrégation de nouveaux facteurs
 - Bureau balance -> aggrégation historique par credit
 - Nouveau facteur:
 - Statut (de 0 à 5) des retards de paiements pondérés par le mois (plus d'importance aux mois récents)
 - STATUS_MW'=-STATUS/(MONTHS_BALANCE-1)
 - Aggregation historique pour chaque credit de la base
 - Nombre :
 - MONTH_
 - Mean et Max
 - STATUS_MW_ et STATUS_
 - Puis jointure
 - avec le fichier Bureau par la clé du crédit: SK_ID_BUREAU
 - Bureau -> aggrégation par client
 - Nouveau facteur:
 - Pourcentage de dette
 - DEBTPCT=AMT_CREDIT_SUM_DEBT/AMT_CREDIT_SUM
 - Aggregation par client et par statut (actif/cloturé):
 - MIN / MAX / MEAN / MEDIANE
 - B_DAYS_APP_, B_DAYS_PD_, B_DAYS_END_,
 - B_PROL_, B_CRED_SUM_, B_CRED_DEBT_, B_CRED_OD_, B_CRED_UPD_,
 - Nouveau facteur : B_CRED_DEBTPCT_
 - Variables issues de bureau balance: BB_MONTH_count_, BB_MONTH_med_, BB_STATUS_mean_, BB_STATUS_max_, BB_STATUS_MW_mean_, BB_STATUS_MW_max_
 - Puis jointure
 - Avec le fichier TRAIN principal par la clé du credit actuel home credit : SK_ID_CURR

4. Creation de facteur

b) Fichiers secondaires

- Aggrégation de nouveaux facteurs
 - **Installments_payments-> aggrégation historique par credit**
 - **Nouveau facteur**
 - retard et différence de paiement
 - $LAG_PAY=DAY_ENTRY_PAYMENT-DAY_INSTALMENT$
 - $DIF_PAY=AMT_PAYMENT-AMT_INSTALMENT$
 - **Aggregation historique pour chaque credit de la base**
 - Nombre et MED:
 - MIN /MAX /MEAN /MED
 - $IP_NUMINST_$
 - MIN /MAX /MEAN /MED
 - $IP_DAYINST_$, $IP_DAYPAY_$
 - $IP_AMTINST_$, $IP_AMTPAY_$
 - $IP_LAGPAY_$, $IP_DIFPAY_$
 - **Puis jointure**
 - avec le fichier PREVIOUS APPLICATION par la clé du crédit: SK_ID_PREV
- **POS_CASH_balance -> aggrégation historique par credit**
 - **Nouveau facteur:**
 - DPD warning et DPD warning pondéré par son age
 - $SK_DPD_WARN=SK_DPD-SK_DPD_DEF$
 - $SK_DPD_WARN_MW= SK_DPD_WARN/(MONTHS_BALANCE-1)$
 - Nombre de paiements restants à la date du CURRENT CREDIT APPLICATION
 - $INST_FUT_APP=MONTHS_BALANCE+CNT_INSTALMENT_FUTURE$
 - **Aggregation historique pour chaque credit de la base**
 - Nombre et MED:
 - MIN / MAX / MEAN /MEDIANE
 - $POS_MONTH_$
 - $POS_INST_$, $POS_INSTFUT_$, $POS_DPD_$, $POS_DPDD_$,
 - Nouveau facteur : $POS_DPDW_$, POS_DPDW_MW , $POS_INSTFUTO_$
 - **Puis jointure**
 - avec le fichier PREVIOUS APPLICATION par la clé du crédit: SK_ID_PREV

4. Creation de facteur

b) Fichiers secondaires

- Aggrégation de nouveaux facteurs
 - **POS_CASH_balance -> aggrégation historique par credit**
 - **Nouveau facteur:**
 - DPD warning
 - **SK_DPD_WARN=SK_DPD-SK_DPD_DEF**
 - **Aggregation historique pour chaque credit de la base**
 - Nombre et MED:
 - **POS_MONTH_**
 - MIN / MAX / MEAN /MEDIANE
 - **CB_LIM_, CB_CDC_ATM_, CB_CDC_, CB_CDC_OTH_, CB_CDC_POS_, CB_DPDD_**
 - **Nouveau facteur : CB_DPDW_**,
 - **Puis jointure**
 - avec le fichier PREVIOUS APPLICATION par la clé du crédit: SK_ID_PREV
 - **Previous_application -> aggrégation par client**
 - **Nouveaux facteurs**
 - TAUX, DUREE, CREDIT/PRIX, APPORT/PRIX,
 - $RATE=(1/CNT_PAYMENT)*(CNT_PAYMENT*AMT_ANNUITY/AMT_CREDIT-1)$
 - $AMT_SUR=(AMT_CREDIT-AMT_APPLICATION)/AMT_CREDIT$
 - $DUR=AMT_CREDIT/AMT_ANNUITY$
 - $CREDPRICE=AMT_CREDIT/AMT_GOODS_PRICE$
 - $APPPRICE=AMT_DOWN_PAYMENT/AMT_GOODS_PRICE$
 - **ONE hot ENCODER**
 - Pour variables categorielles
 - **NAME_CONTRACT_TYPE, NAME_CONTRACT_STATUS, NAME_PORTFOLIO, NAME_YIELD_GROUP**
 - **Aggregation par client**
 - **Puis jointure**
 - Avec le fichier TRAIN principal par la clé du credit actuel home credit : SK_ID_CURR

5. MODELISATION

A) PREPROCESSING

REEQUILIBRAGE DES CLASSES TARGET PAR UNDERSAMPLING ALEATOIRE DE LA TABLE TRAIN

Oversampling possible mais nous sommes preneur d'une reduction des échantillons pour accelerer les calculs

SELECTION DE VARIABLES A TESTER

Une réduction d'univers grâce à l'analyse exploratoire, c'est toujours ça de gagné

SPLIT de TRAIN EN TRAIN 70%/TEST 30%

En gardant la même proportion des classes dans les 2 échantillons

FIT One Hot Encoder sur les variables catégorielles de XTRAIN

TRANSFORM One Hot Encoder sur les variables catégorielles de XTRAIN

TRANSFORM One Hot Encoder sur les variables catégorielles de XTEST

FIT MINMAX scaler sur XTRAIN

TRANSFORM MINMAX scaler sur XTRAIN

TRANSFORM MINMAX scaler sur XTEST

FIT IMPUTATION KNN sur XTRAIN

TRANSFORM IMPUTATION KNN sur XTRAIN

TRANSFORM IMPUTATION KNN sur XTEST

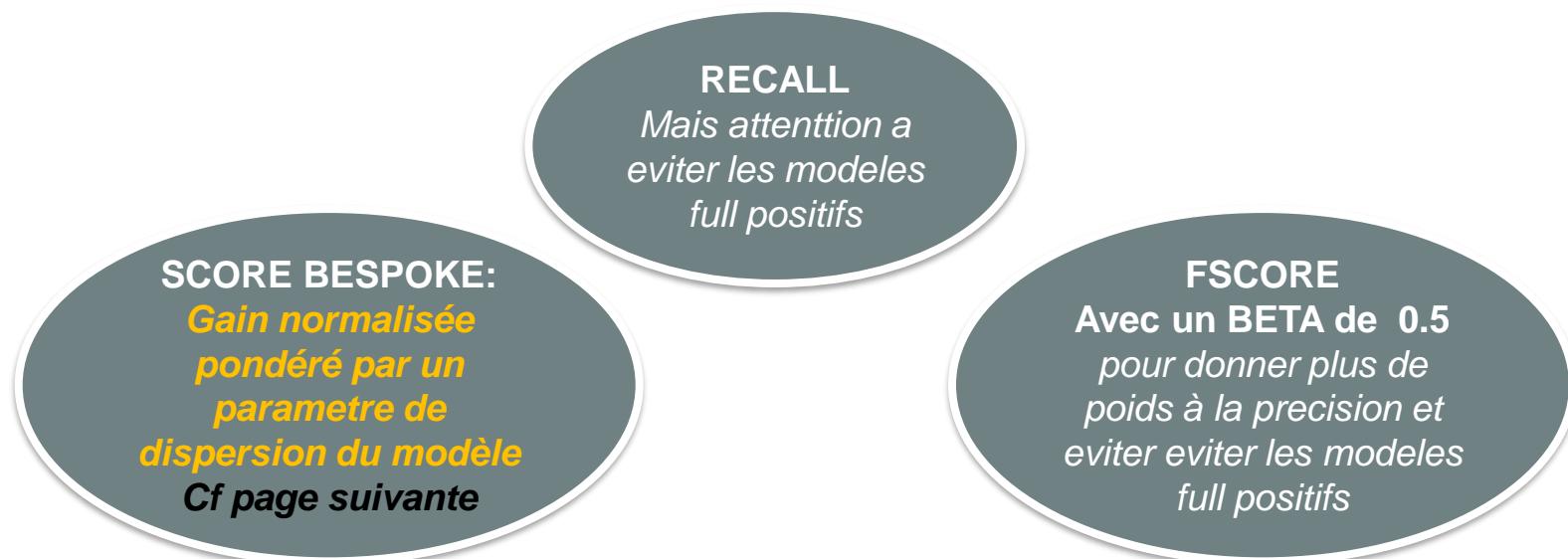
XTRAIN/YTRAIN
Prêt pour ENTRAINEMENT ET VALIDATION CROISEE

XTEST/YTEST
Prêt pour TESTING

5. MODELISATION

B) Selection et creation de fonction de scoring pour les gridsearch cross validation

- Les défauts étant plus couteux que les rentrées d'intérêt surtout après équilibrage des classes, les FN à minimiser sont plus importants que les FP à minimiser
- Les FN sont à minimiser pour éviter les défauts très couteux
 - **Donc TP/P réel ou RECALL classe positive à maximiser**
 - Sans rentrer dans le piège du modèle full positif (tout le monde est prédict défaillant, pas de crédit octroyé, pas de prise de risque) qui donnerait un recall de 1
 - à pondérer dans ce cas avec la précision: le modèle full positif donnerait une précision de 0,5 dans notre cas de classes équilibrées
- Les FP sont à minimiser pour maximiser les rentrées
 - **Donc TN/N réel ou RECALL classe négative à maximiser**



5. MODELISATION

B) Selection et creation de fonction de scoring pour les gridsearch cross validation

Gain maximum possible dans la réalité:

$$\sum_{i=0}^n i_{NEGATIFS} X i_{T-INTERETS} X i_{CREDIT}$$

Gain maximum possible MODELISATION:

$$\sum_{i=0}^n i_{TN \text{ ou } FP} X i_{T-INTERETS} X i_{CREDIT}$$

Proche de 0.5 x 0.02
Si classes équilibrées , taux de 2% et crédit normalise à 1

Cas sans prise de risque (tout positif):
0 finalement proche du max

Perte maximale possible dans la réalité:
(hypothèse d'un défaut total)

$$-\sum_{i=0}^n i_{POSITIFS} X i_{CREDIT}$$

Perte maximale possible : MODELISATION

$$-\sum_{i=0}^n i_{TP \text{ ou } FN} X i_{CREDIT}$$

Proche de -0.5
Si classes équilibrées et crédit normalise à 1

Cas avec prise de risque max (tout negatif):
Proche de -0.5 finalement proche du max

- DANS NOTRE CAS : CLASSES EQUILIBREES / FN COUTEUX / TN REMUNERATEUR
- FONCTION DE SCORING BESPOKE NORMEE ENTRE 0 ET 1 A MAXIMISER:
 - ONDREEE lineairement (sinh utilisable) PAR UN PARAMETRE DE DISPERSION PENALISANT LES PREDICTIONS TROP HOMOGENES (une des classes trop majoritaire)

$$\sqrt{\frac{(\sum_{i=0}^n i_{FP} \times i_{T-INTERETS} \times i_{CREDIT} + \sum_{i=0}^n i_{TP} X i_{CREDIT}) \times (1 - 2 \times |0.5 - \frac{TP + FP}{TP + FP + TN + FN}|)}{(\sum_{i=0}^n i_{TN \text{ ou } FP} \times i_{T-INTERETS} \times i_{CREDIT} + \sum_{i=0}^n i_{FN \text{ ou } FN} X i_{CREDIT})}}$$

5. MODELISATION

C) MODELISATION - ALGORITHMES

*REGRESSION
LOGISTIQUE*

*LINEAR SVM
CLASSIFIEUR*

*SVM
CLASSIFIEUR*

BAGGING

ADABOOST

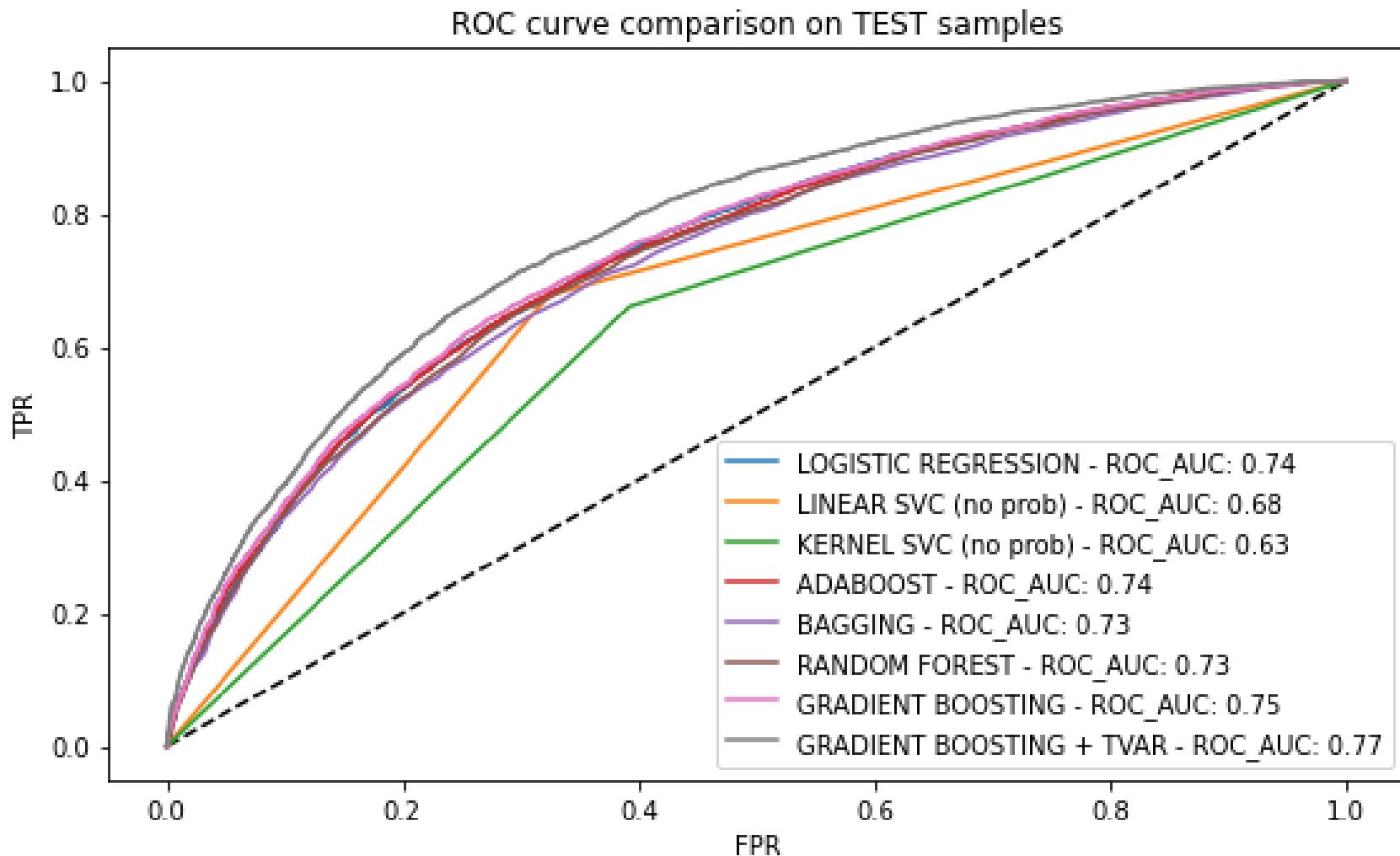
*RANDOM
FOREST*

*GRADIENT
BOOSTING*

RECHERCHE D'HYPER PARAMETRES AVEC GRIDSEARCHCV

5. MODELISATION

C) MODELISATION – ALGORITHMES



5. MODELISATION

C) MODELISATION – ALGORITHMES

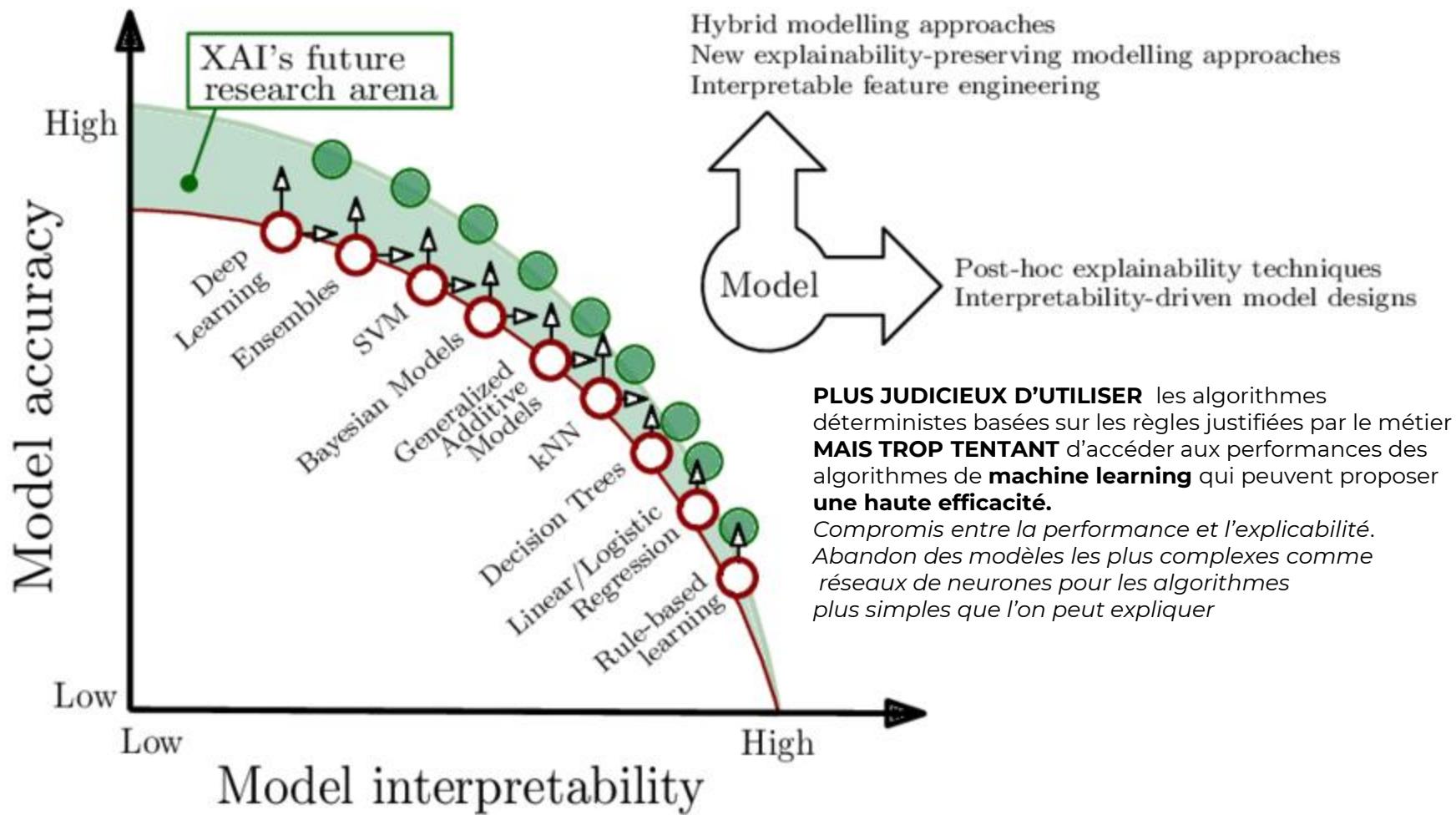
	Test ROC_AUC	Test Accuracy	Test recall 0	Test recall 1	Test precision 0	Test precision 1	Test fscore 0	Test fscore 1	GSCV fit time MIN	GSCV fit time MAX	GSCV HYPERPARAM 1	GSCV HYPERPARAM 2	GSCV KFOLD
LOGISTIC REGRESSION	0.74	0.68	0.67	0.68	0.68	0.68	0.68	0.68	0.15	2.29	C	I1_ratio	5
LINEAR SVC (no prob)	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.08	5.43	C	NaN	5
KERNEL SVC (no prob)	0.63	0.63	0.61	0.66	0.64	0.63	0.62	0.64	20.79	96.08	C	kernel & gamma	3
ADABOOST	0.74	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.52	10.55	n_estimators	learning_rate	5
BAGGING	0.73	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.24	66.94	n_estimators	NaN	5
RANDOM FOREST	0.73	0.68	0.67	0.68	0.68	0.67	0.67	0.68	0.21	68.27	n_estimators	max_features	5
GRADIENT BOOSTING	0.75	0.68	0.68	0.69	0.68	0.68	0.68	0.68	2.25	39.28	n_estimators	max_leaf_nodes	5
GRADIENT BOOSTING + TVAR	0.77	0.70	0.58	0.82	0.76	0.66	0.66	0.73	16.66	290.60	n_estimators	max_leaf_nodes	5

Apport des variables transformées

Voir en annexes la recherche d'hyperparamètres

5. MODELISATION

D) INTERPRETABILITE



5. MODELISATION

D) - 1 INTERPRETABILITE Globale – GRADIENT BOOSTING

L'importance des facteurs/caractéristiques basées sur les impuretés de GINI

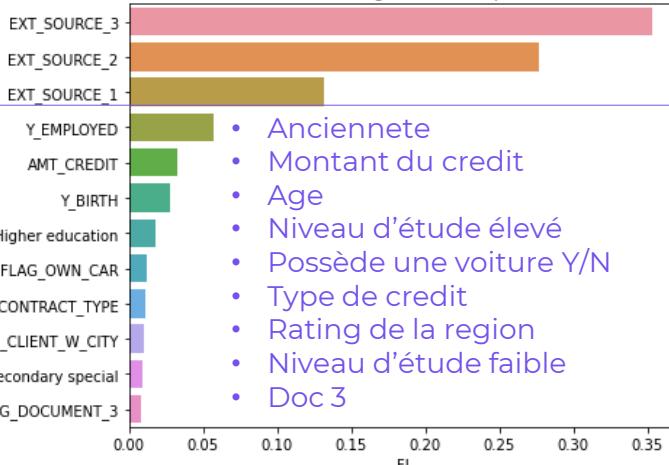
L'importance d'une caractéristique est calculée comme la réduction totale (normalisée) du critère apporté par cette caractéristique

les importances des caractéristiques basées sur les impuretés peuvent être trompeuses (élevées) pour les caractéristiques à cardinalité élevée (nombreuses valeurs uniques) qui ne sont pas forcément prédictives de la variable cible

- Les scores externes sont très importants

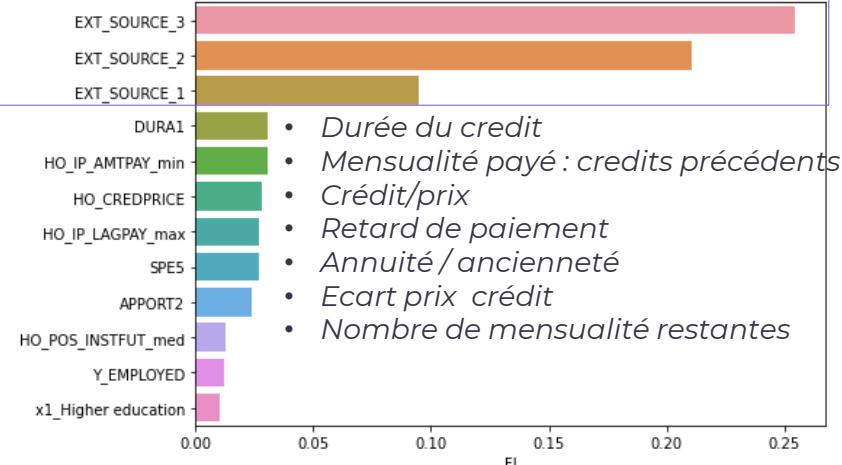


Gradient Boosting Features Importances



Modèle avec les variables du fichier data_train

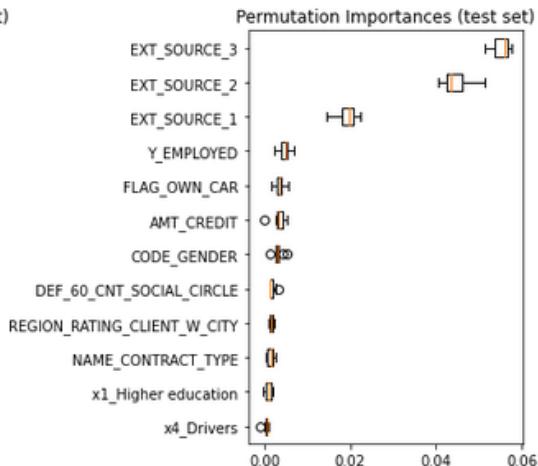
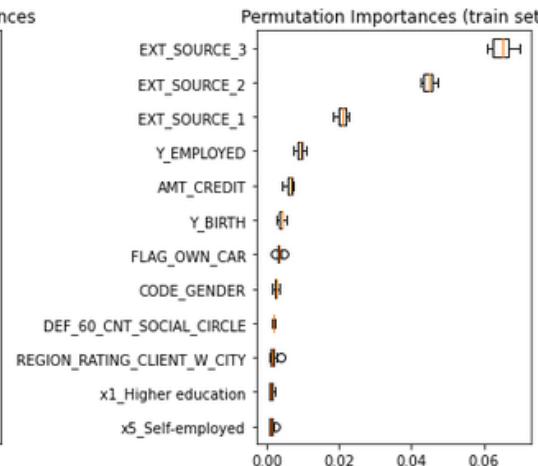
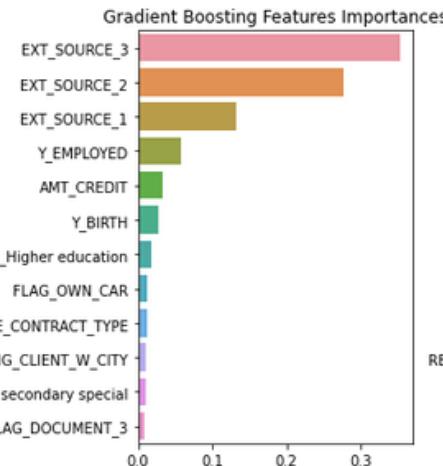
Gradient Boosting Features Importances



Modèle avec les variables du fichier data_train + variables transformées et/ou issues des autres tables

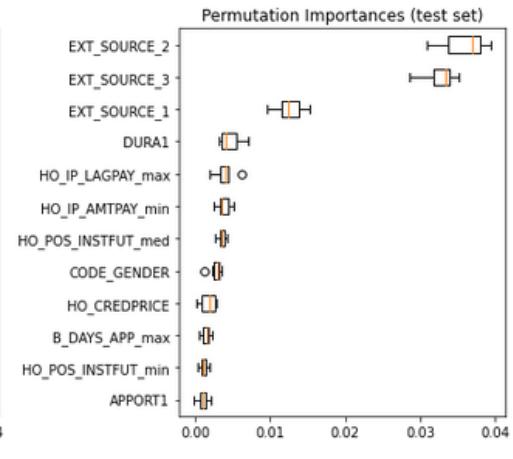
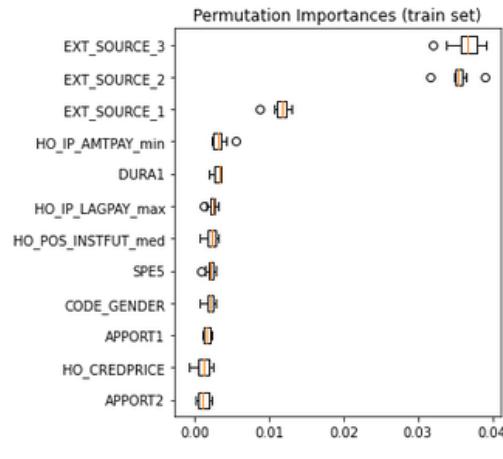
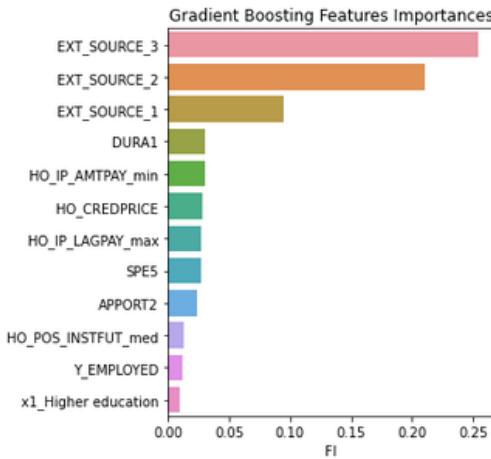
5. MODELISATION

D) - 1 INTERPRETABILITE Globale – GRADIENT BOOSTING



Modèle avec les variables du fichier data_train

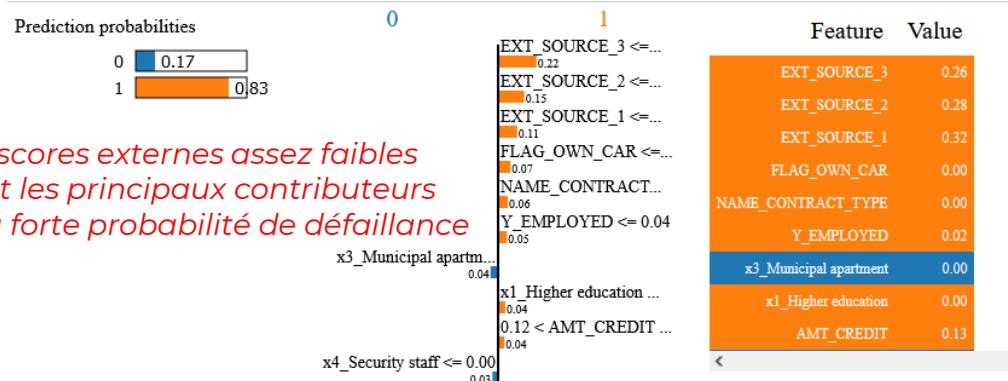
Modèle avec les variables du fichier data_train + variables transformées et/ou issues des autres tables



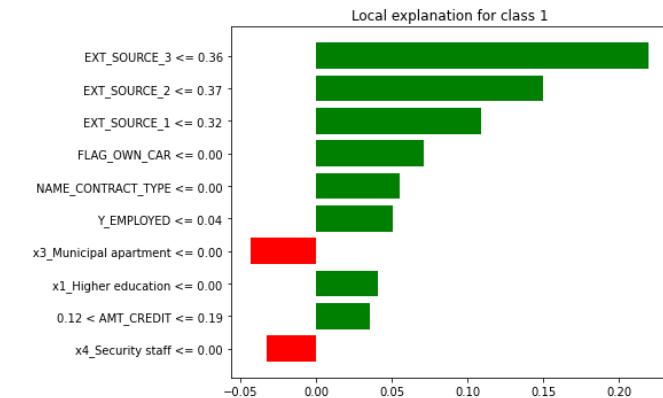
5. MODELISATION

D) - 1 INTERPRETABILITE Locale avec LIME – GRADIENT BOOSTING : *individu qui a une probabilité de 85% d etre defaillant*

Modèle avec les variables du fichier data_train

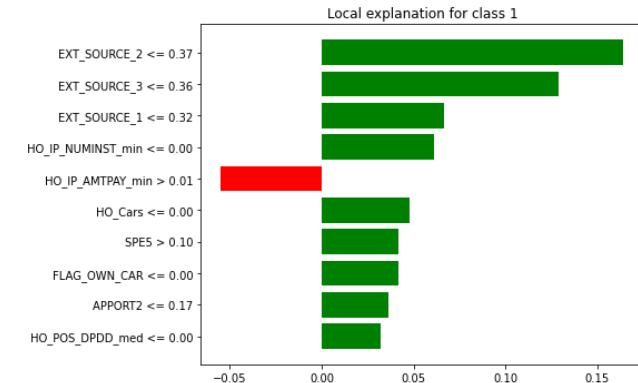
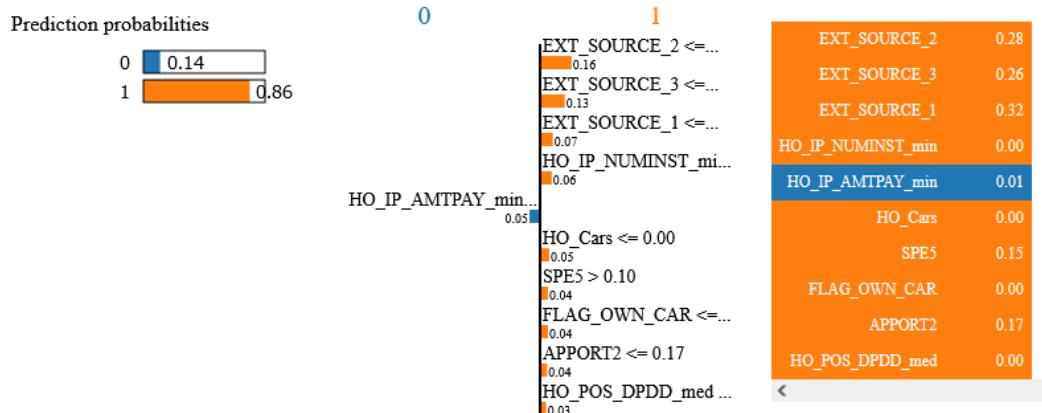


Les scores externes assez faibles
sont les principaux contributeurs
à la forte probabilité de défaillance



Les valeurs sont celles après preprocessing , pour bien faire,
il faudrait remonter les valeurs initiales

Modèle avec les variables du fichier data_train + variables transformées et/ou issues des autres tables



D1

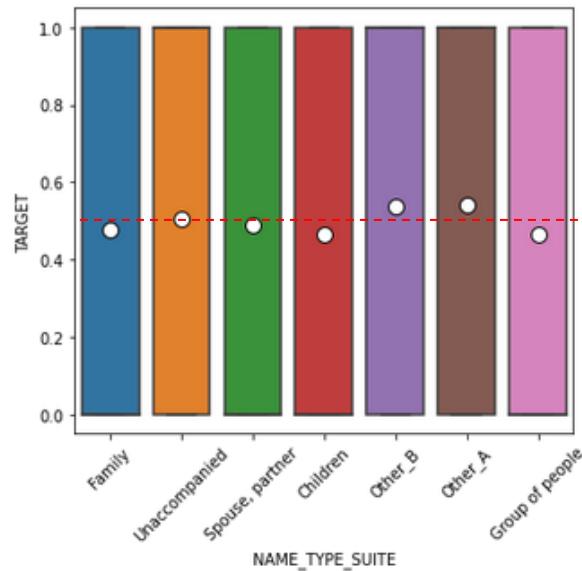
ANNEXES ANALYSE DESCRIPTIVE CATEGORIES

LES ENCADREMENTS SIGNALENT DES MODALITES SUFFISAMMENT REPRESENTEES

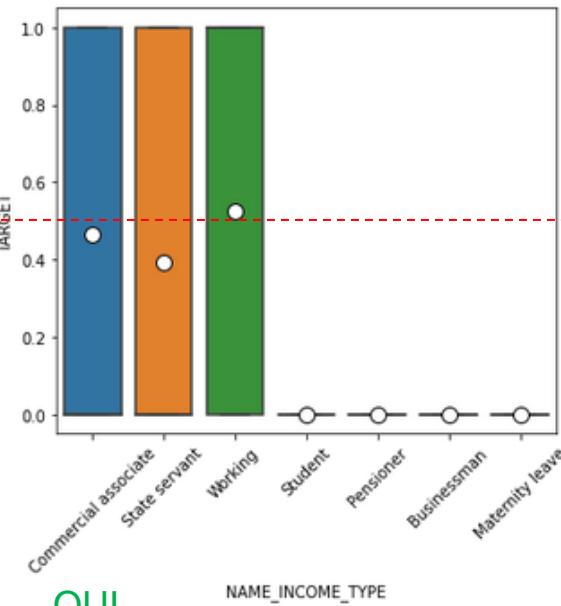
Unaccompanied	325741
Family	47794
Spouse, partner	14688
Children	3366
Other_B	2386
Other_A	1222
Group of people	311
Name: NAME_TYPE_SUITE, dtype: int64	

Working	259795
Commercial associate	105601
State servant	30086
Student	14
Pensioner	7
Businessman	4
Maternity leave	1
Name: NAME_INCOME_TYPE, dtype: int64	

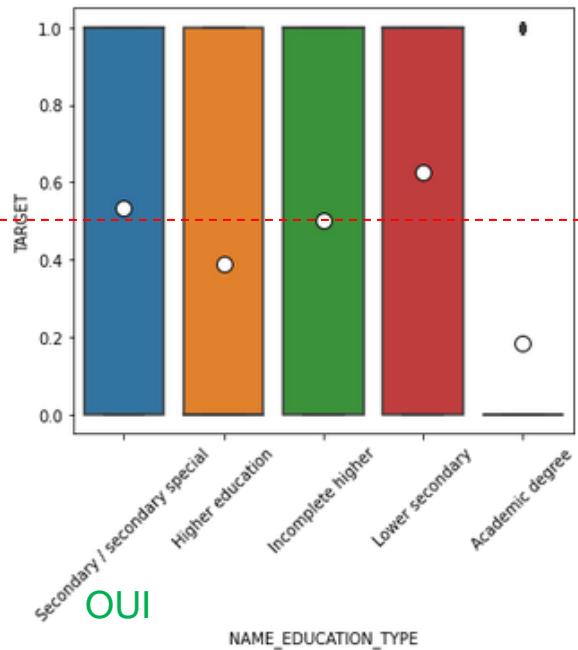
Secondary / secondary special	286188
Higher education	90031
Incomplete higher	14772
Lower secondary	4371
Academic degree	146
Name: NAME_EDUCATION_TYPE, dtype: int64	



NON



OUI



OUI

LES ENCADREMENTS SIGNALENT DES MODALITES SUFFISAMMENT REPRESENTEES

Married	253989
Single / not married	63266
Civil marriage	43359
Separated	25700
Widow	9194

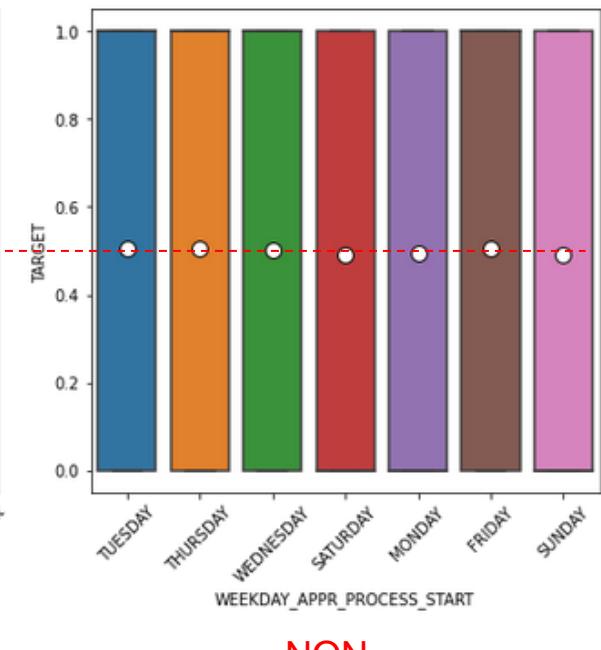
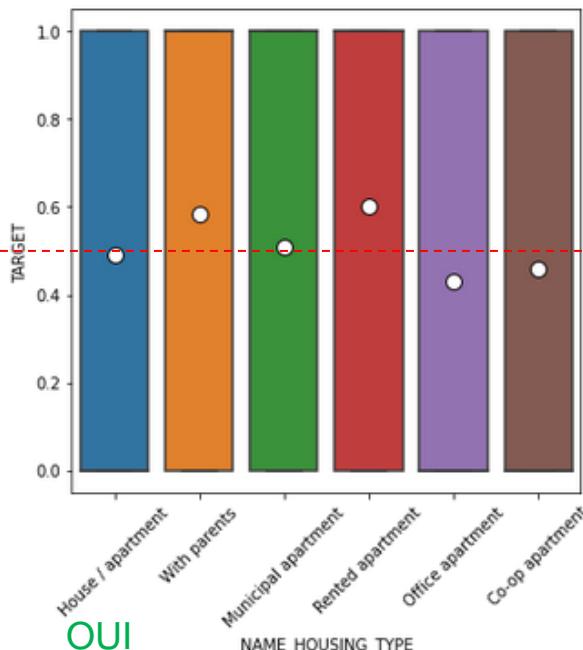
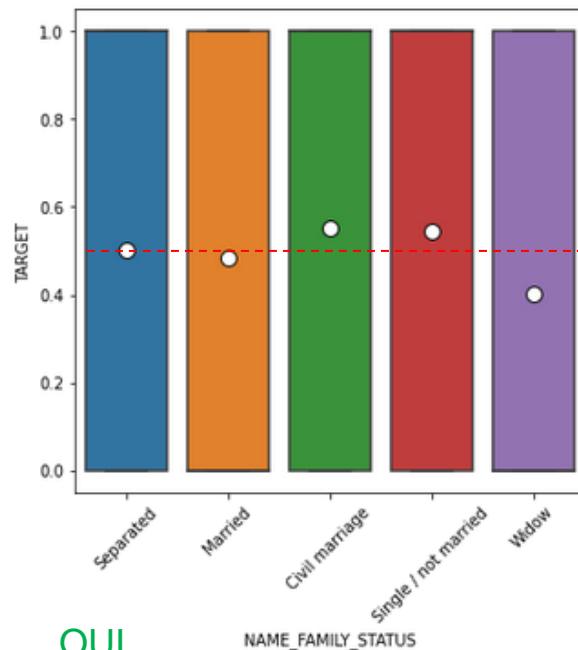
Name: NAME_FAMILY_STATUS, dtype:

House / apartment	343068
With parents	25048
Municipal apartment	14388
Rented apartment	8245
Office apartment	3274
Co-op apartment	1485

Name: NAME_HOUSING_TYPE, dtype:

TUESDAY	67830
WEDNESDAY	67733
FRIDAY	65423
THURSDAY	65279
MONDAY	63709
SATURDAY	44442
SUNDAY	21092

Name: WEEKDAY_APPR_PROCESS_START



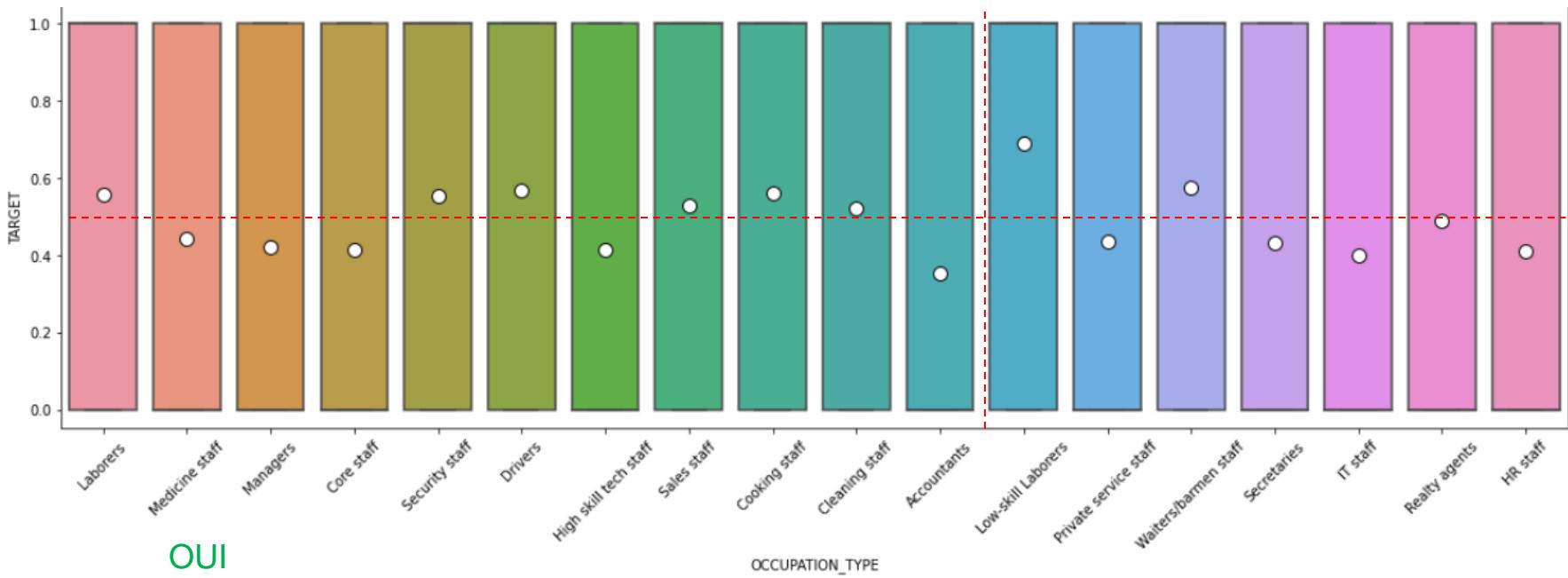
On peut identifier un comportement de valeur si la moyenne dans nos boxplot s'écartent de 0,5

Laborers	95006
Sales staff	51199
Core staff	38599
Drivers	32556
Managers	30525
High skill tech staff	15919
Medicine staff	12804
Accountants	12751
Security staff	11374
Cooking staff	10329
Cleaning staff	7437

LES ENCADREMENTS SIGNALENT DES MODALITES SUFFISAMMENT REPRESENTEES

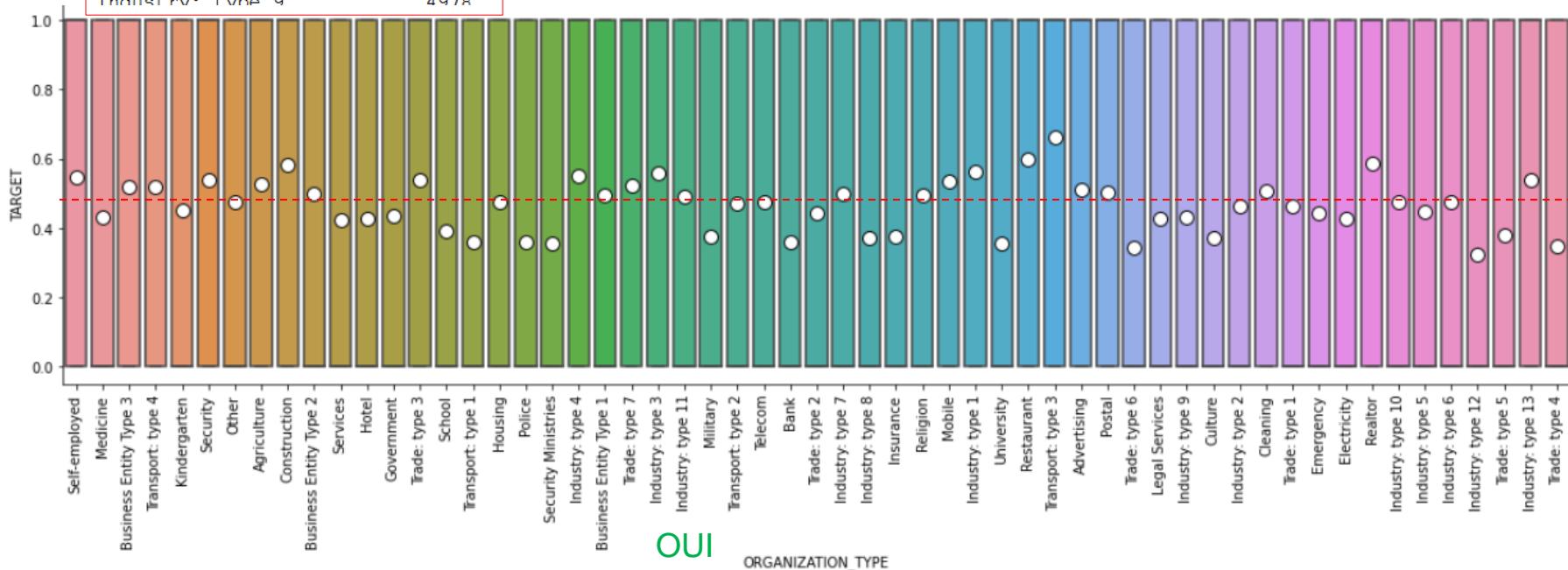
Low-skill Laborers	4293
Private service staff	3652
Waiters/barmen staff	2282
Secretaries	1852
Realty agents	1120
HR staff	795
IT staff	704

Name: OCCUPATION_TYPE, dtype: int64



Business Entity Type 3	109335
Self-employed	62748
Other	25241
Business Entity Type 2	17124
Medicine	16404
Government	15180
Trade: type 7	12361
School	12268
Construction	11948
Kindergarten	10315
Business Entity Type 1	9457
Transport: type 4	8777
Trade: type 3	5728
Industry: type 3	5704
Security	5445
Industry: type 9	4978

LES ENCADREMENTS SIGNALENT DES MODALITES SUFFISAMMENT REPRESENTEES



On peut identifier un comportement de valeur si la moyenne dans nos boxplot s'écartent de 0,5

D2

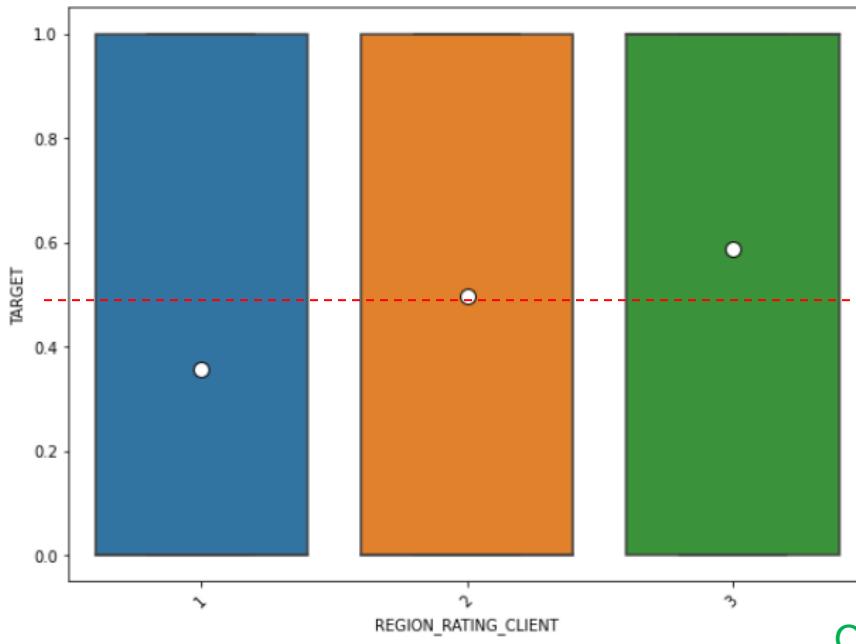
ANNEXES ANALYSE DESCRIPTIVE QUANTITATIVES ORDINALES

• Les ratings clients apportent de la valeur

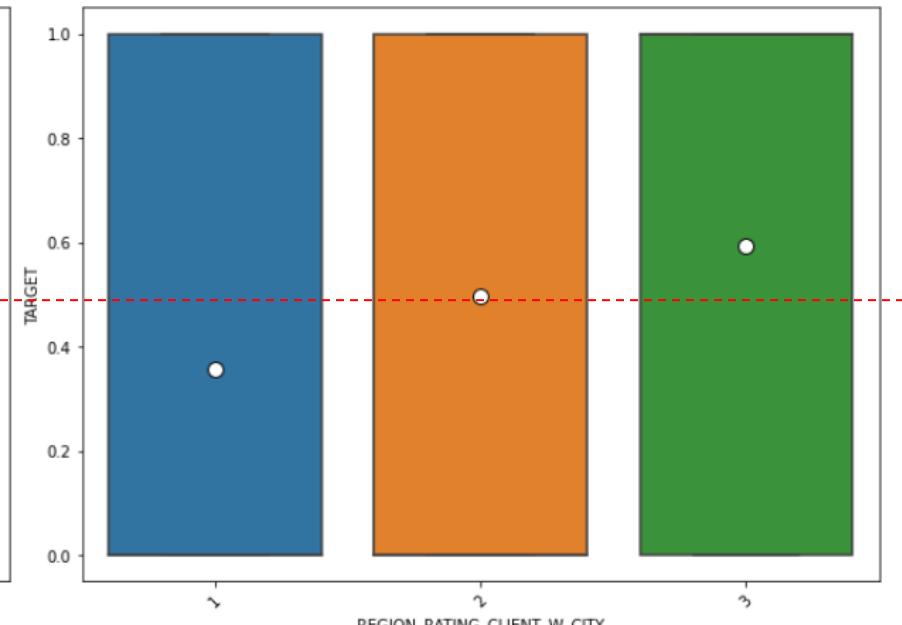
```
2    290290
3    70866
1    34352
Name: REGION_RATING_CLIENT, dtype: int64
2    293755
3    65187
1    36566
Name: REGION_RATING_CLIENT_W_CITY, dtype: int64
```

Corrélation de rang de Spearman:
Correlation des ratings avec la target

	TARGET	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
TARGET	1.000000	0.109560	0.114296
REGION_RATING_CLIENT	0.109560	1.000000	0.952755
REGION_RATING_CLIENT_W_CITY	0.114296	0.952755	1.000000



OUI

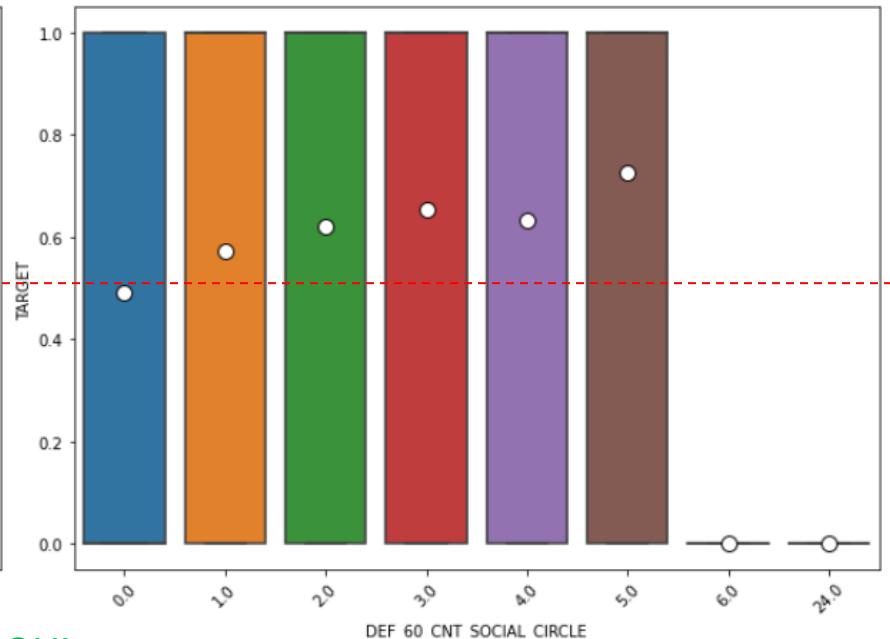
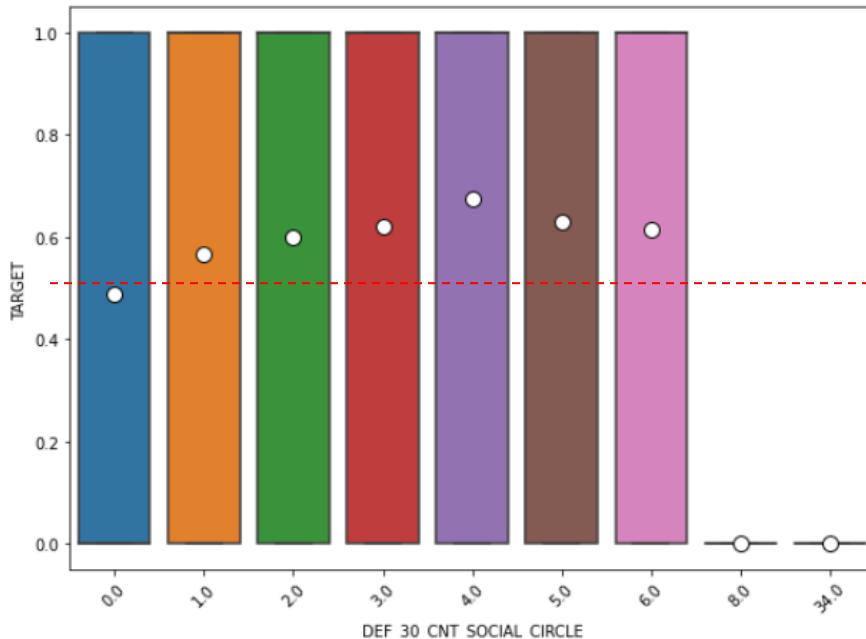


• Le contexte social apporte de la valeur

0.0	344851
1.0	40245
2.0	8040
3.0	1856
4.0	404
5.0	97
6.0	13
8.0	1
34.0	1
Name: DEF_30_CNT_SOCIAL_CIRCLE,	

0.0	358084
1.0	31195
2.0	4992
3.0	987
4.0	204
5.0	44
6.0	1
24.0	1
Name: DEF_60_CNT_SOCIAL_CIRCLE,	

*LES ENCADREMENTS ROUGES
SIGNALENT DES MODALITES
SUFFISAMMENT REPRESENTEES*

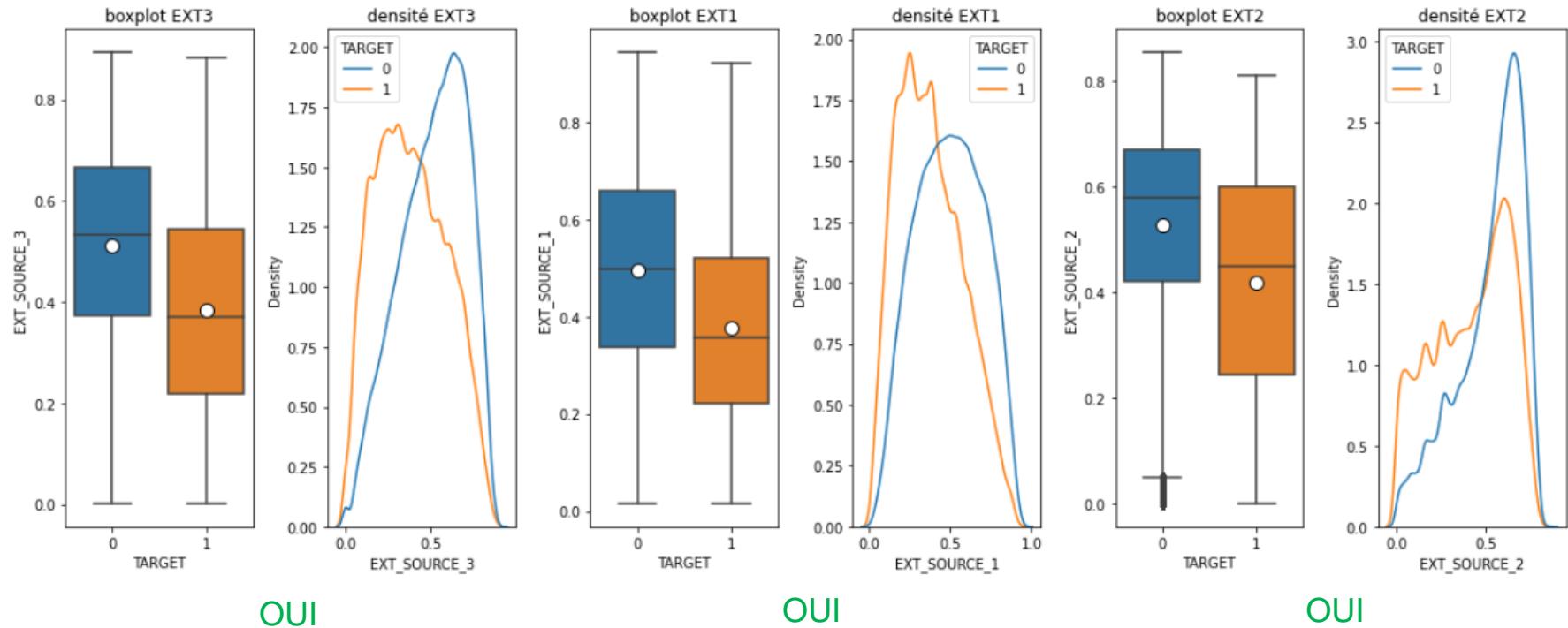


OUI

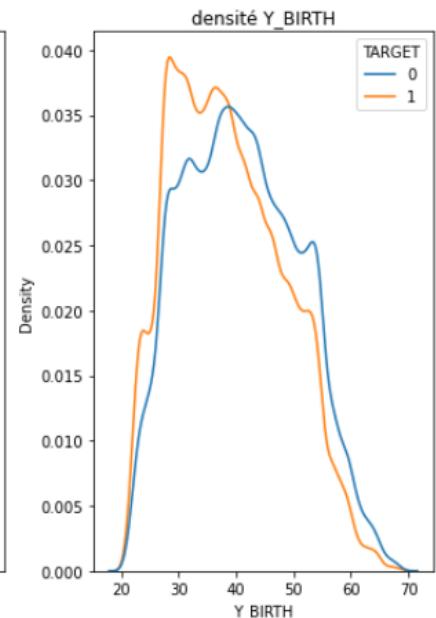
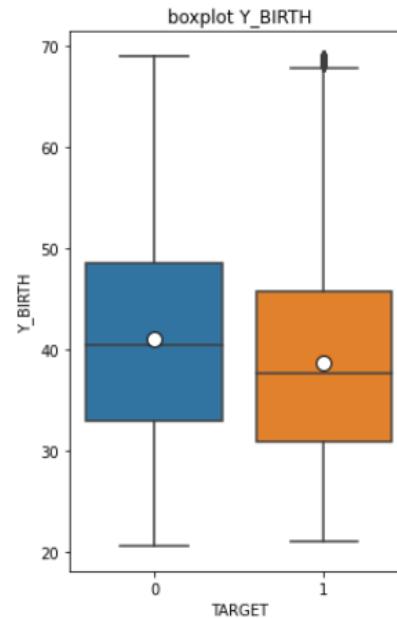
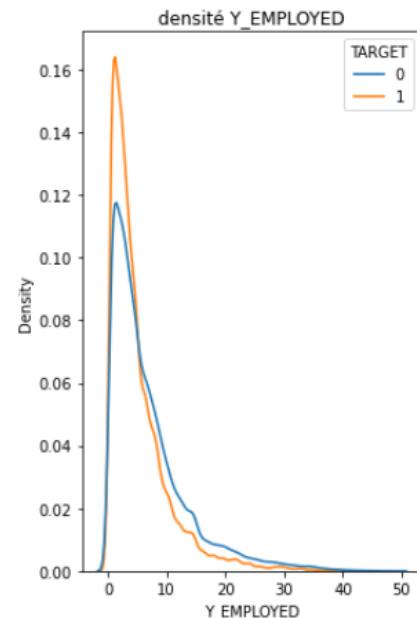
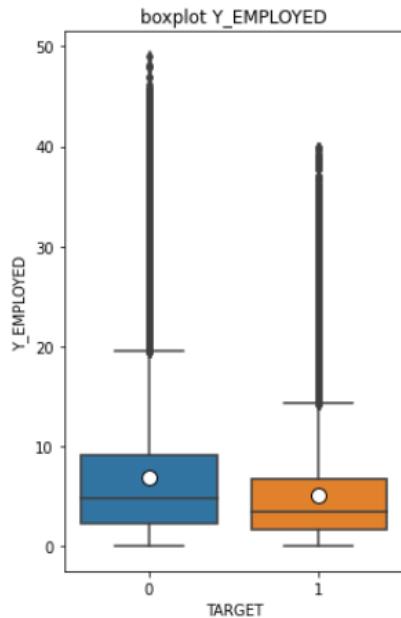
D3

ANNEXES ANALYSE DESCRIPTIVE QUANTITATIVES CONTINUES

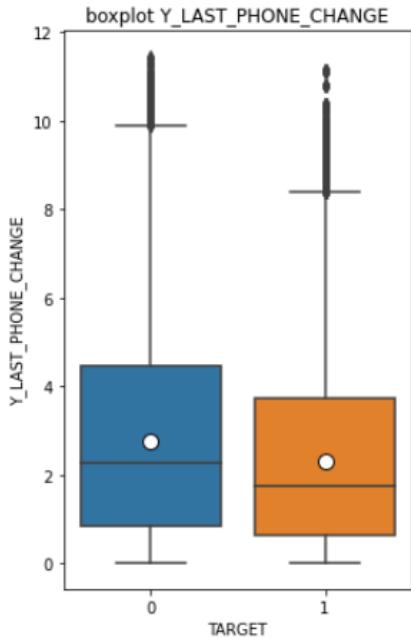
- **Les scores externes apportent de la valeur**



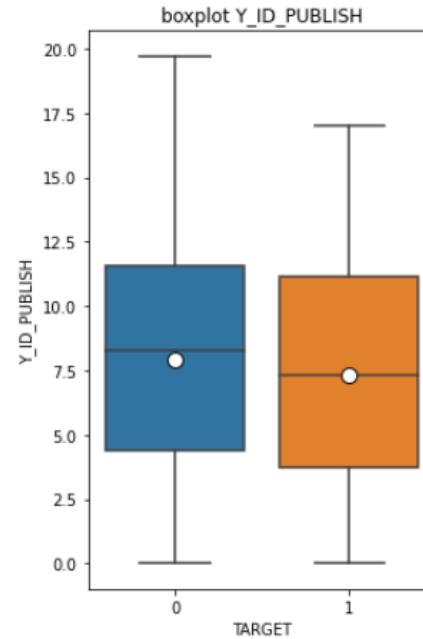
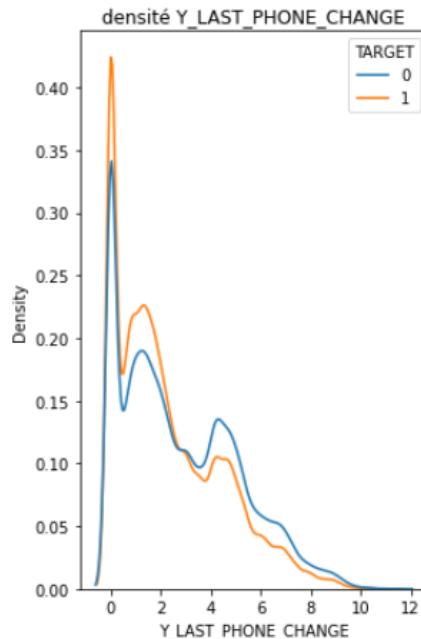
- **Les ages et durée du dernier emploi fonctionnent**
 - Les ages faibles et courtes durées d'emploi augmentent le risque de défaillance



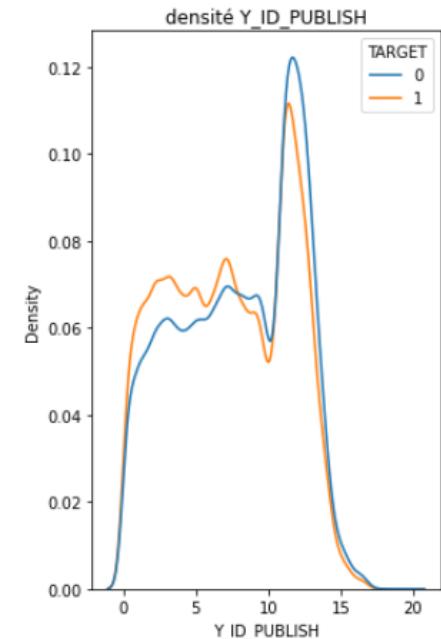
- Les dates de changement de téléphone et de papiers d'identité : PEUT ETRE?**



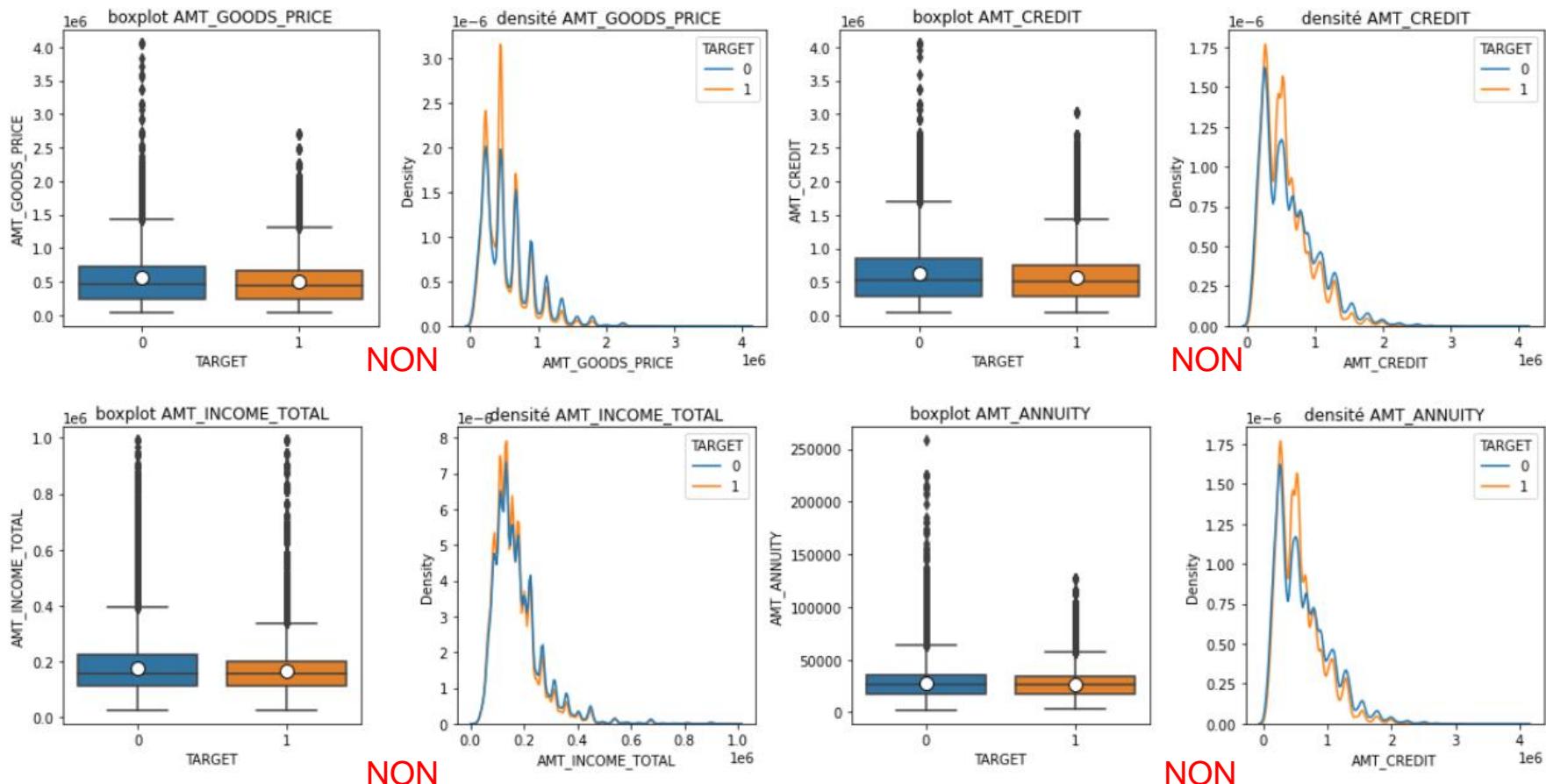
PEUT
ETRE



PEUT
ETRE



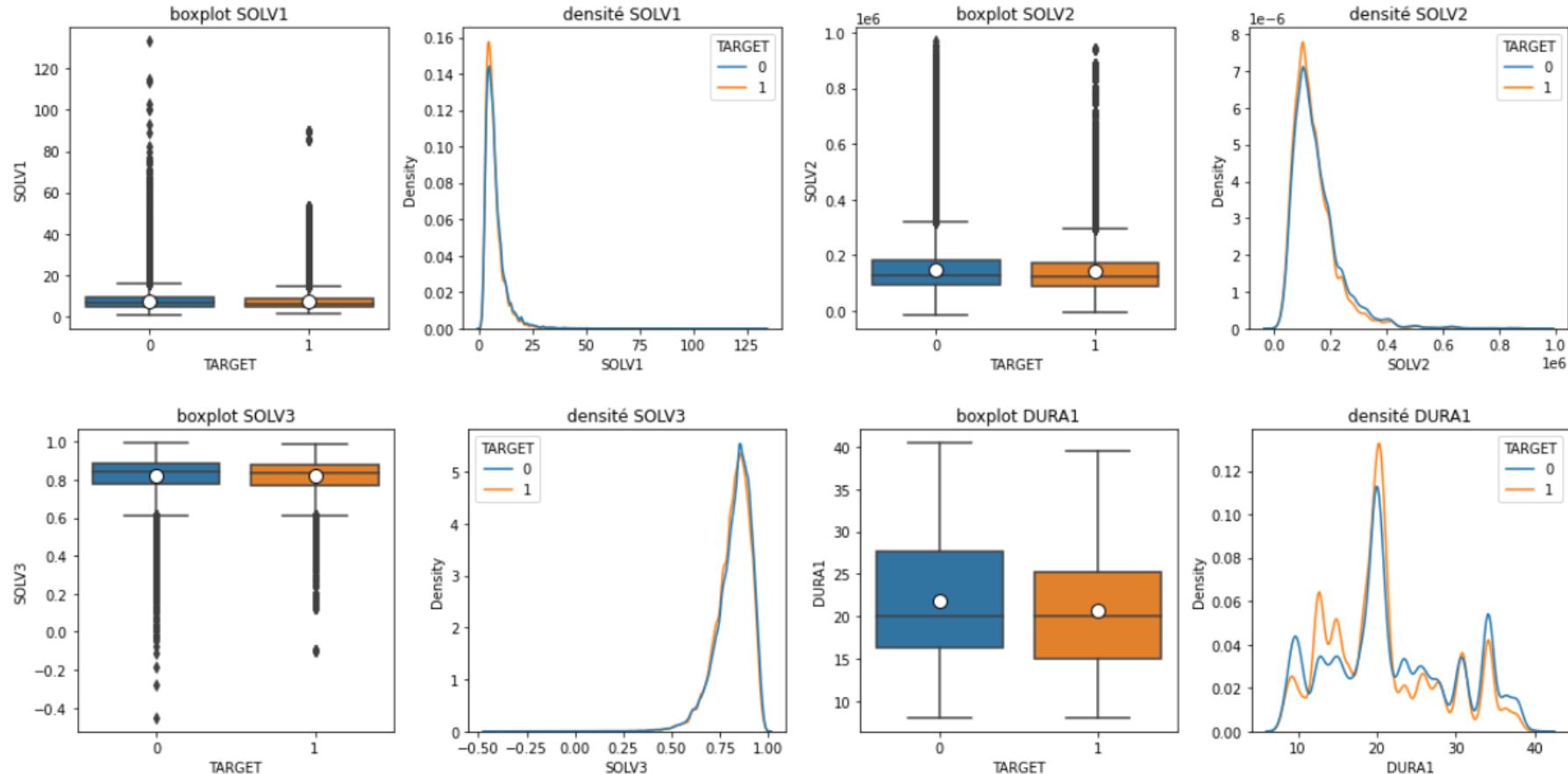
- **Les montants des credits, annuité , revenus et des biens pas vraiment discriminant**



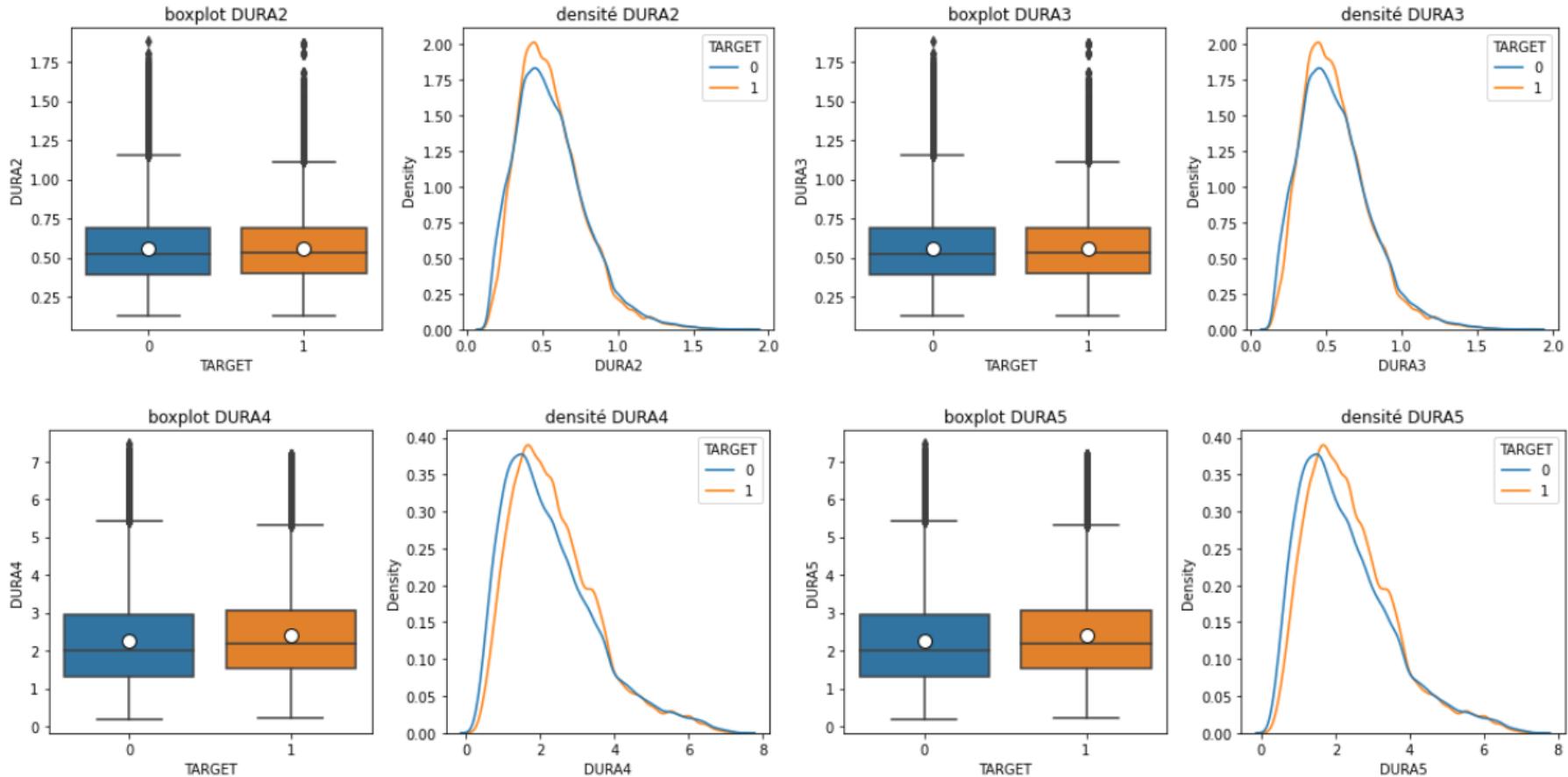
D4

ANNEXES ANALYSE DESCRIPTIVE QUANTITATIVES CONTINUES FACTEURS CREEES DEPUIS FICHIER APP TRAIN

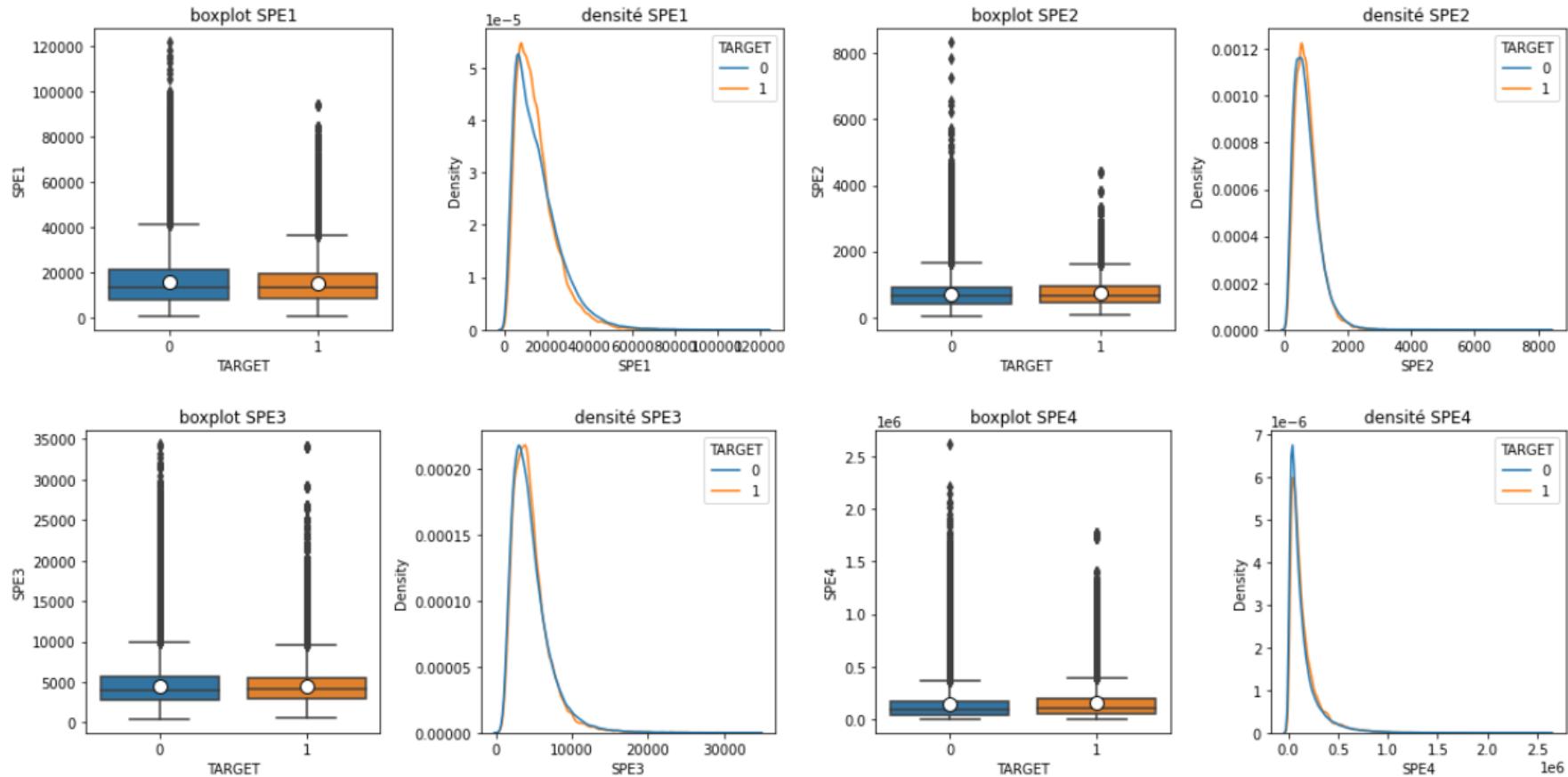
Création de nouveaux facteurs à partir du fichier principal:



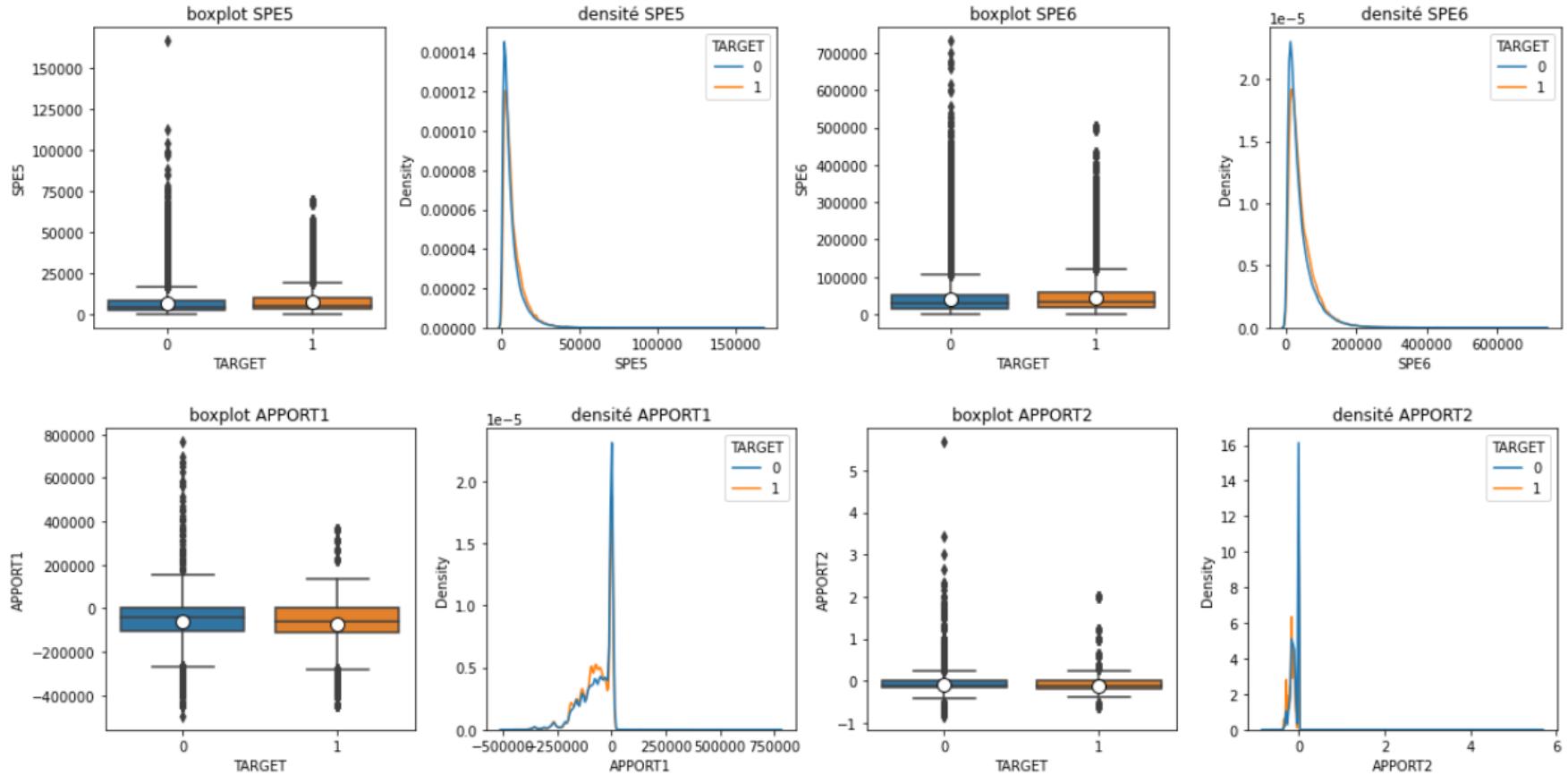
Création de nouveaux facteurs à partir du fichier principal:



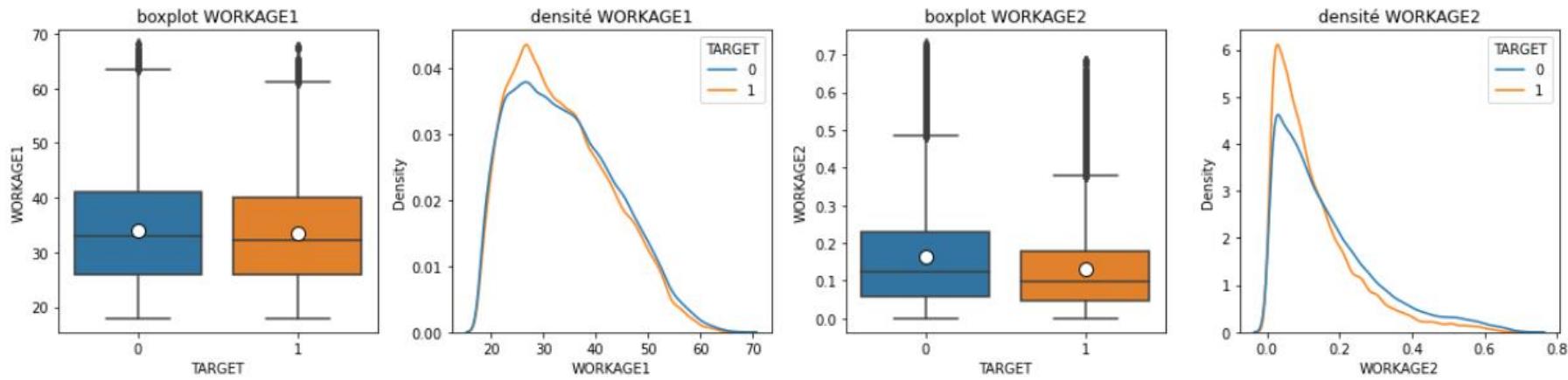
Création de nouveaux facteurs à partir du fichier principal:



Création de nouveaux facteurs à partir du fichier principal:



Création de nouveaux facteurs à partir du fichier principal:



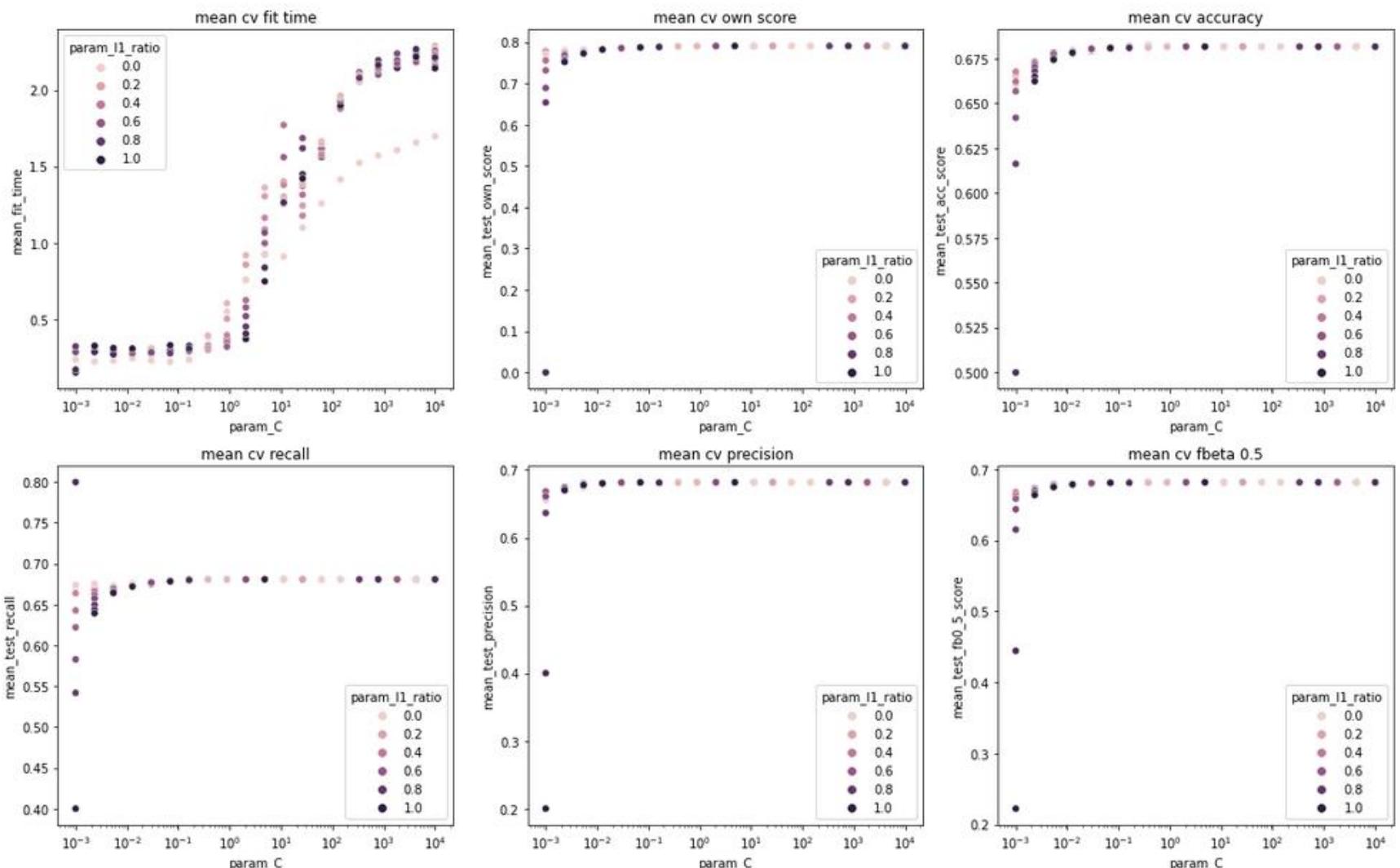
D5

ANNEXES GRIDSEARCHCV

```

meilleur estimateur LogisticRegression(C=2.069138081114788, l1_ratio=1.0, max_iter=500,
penalty='elasticnet', solver='saga', tol=0.001)
best score 0.79
best param {'C': 2.069138081114788, 'dual': False, 'l1_ratio': 1.0, 'max_iter': 500, 'penalty': 'elasticnet', 'solver': 'sa
ga', 'tol': 0.001}
score prediction 0.79
roc auc: 0.74

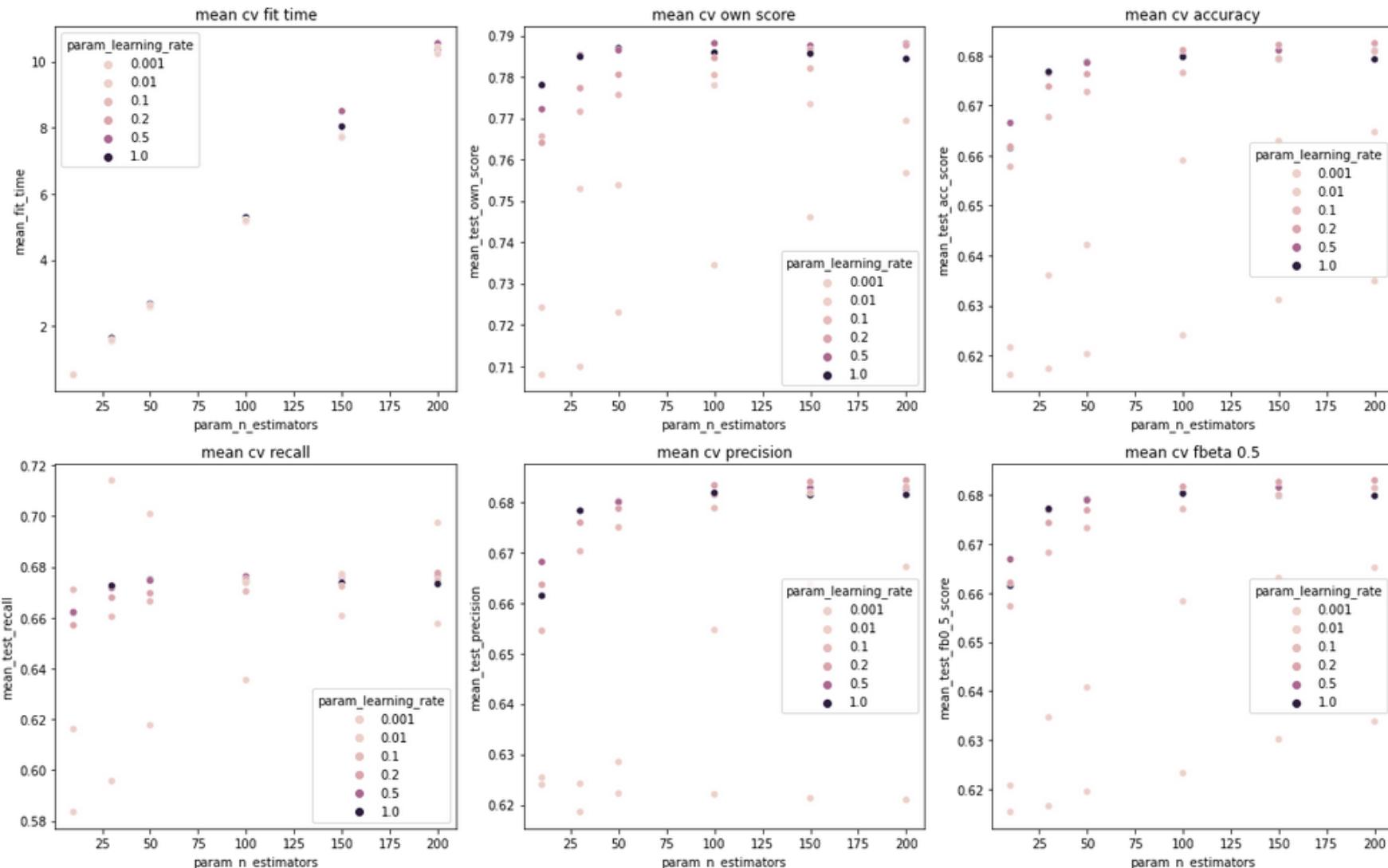
```



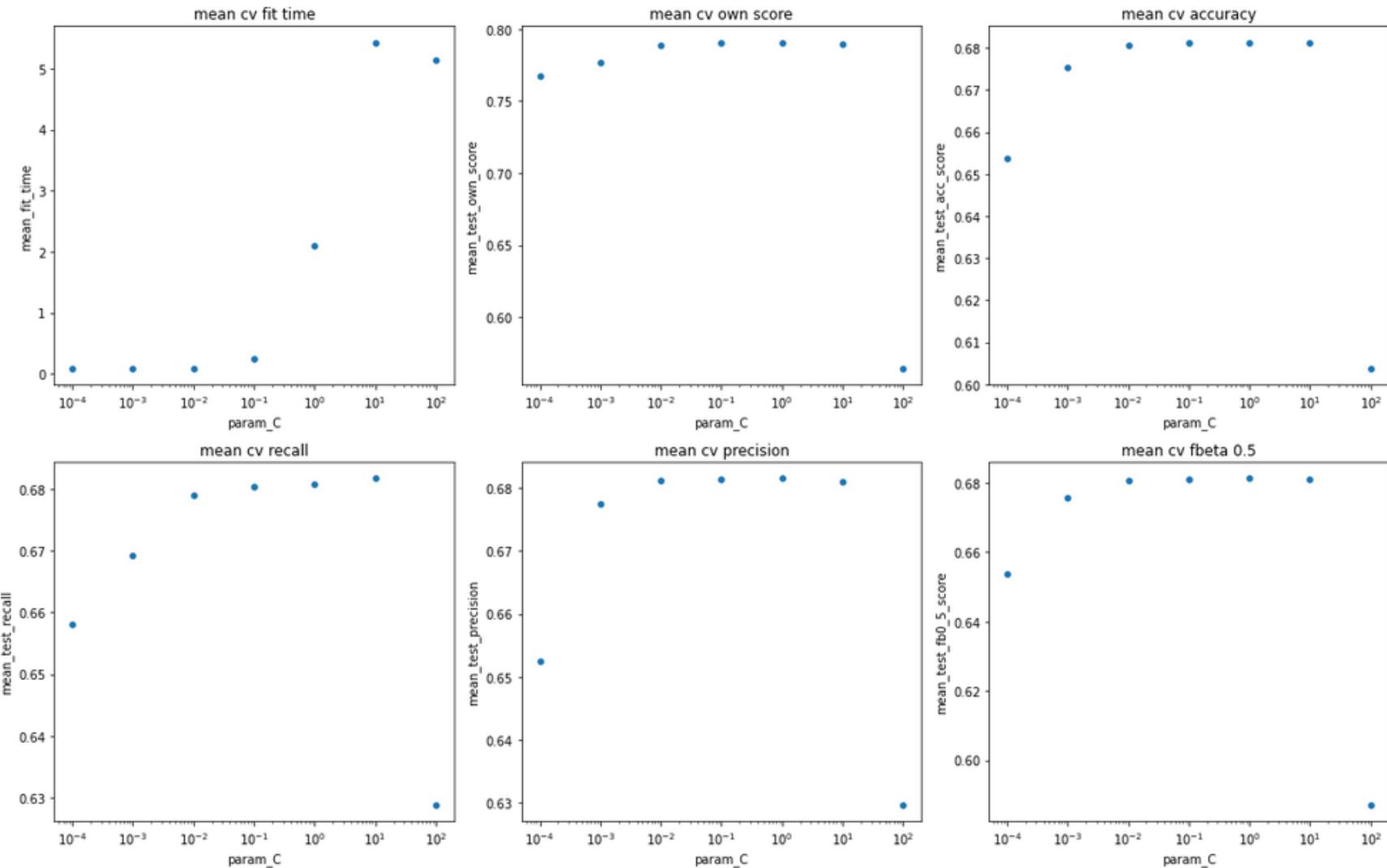
```

meilleur estimateur AdaBoostClassifier(learning_rate=0.5, n_estimators=100)
best score 0.79
best param {'learning_rate': 0.5, 'n_estimators': 100}
score prediction 0.79
roc auc: 0.74

```



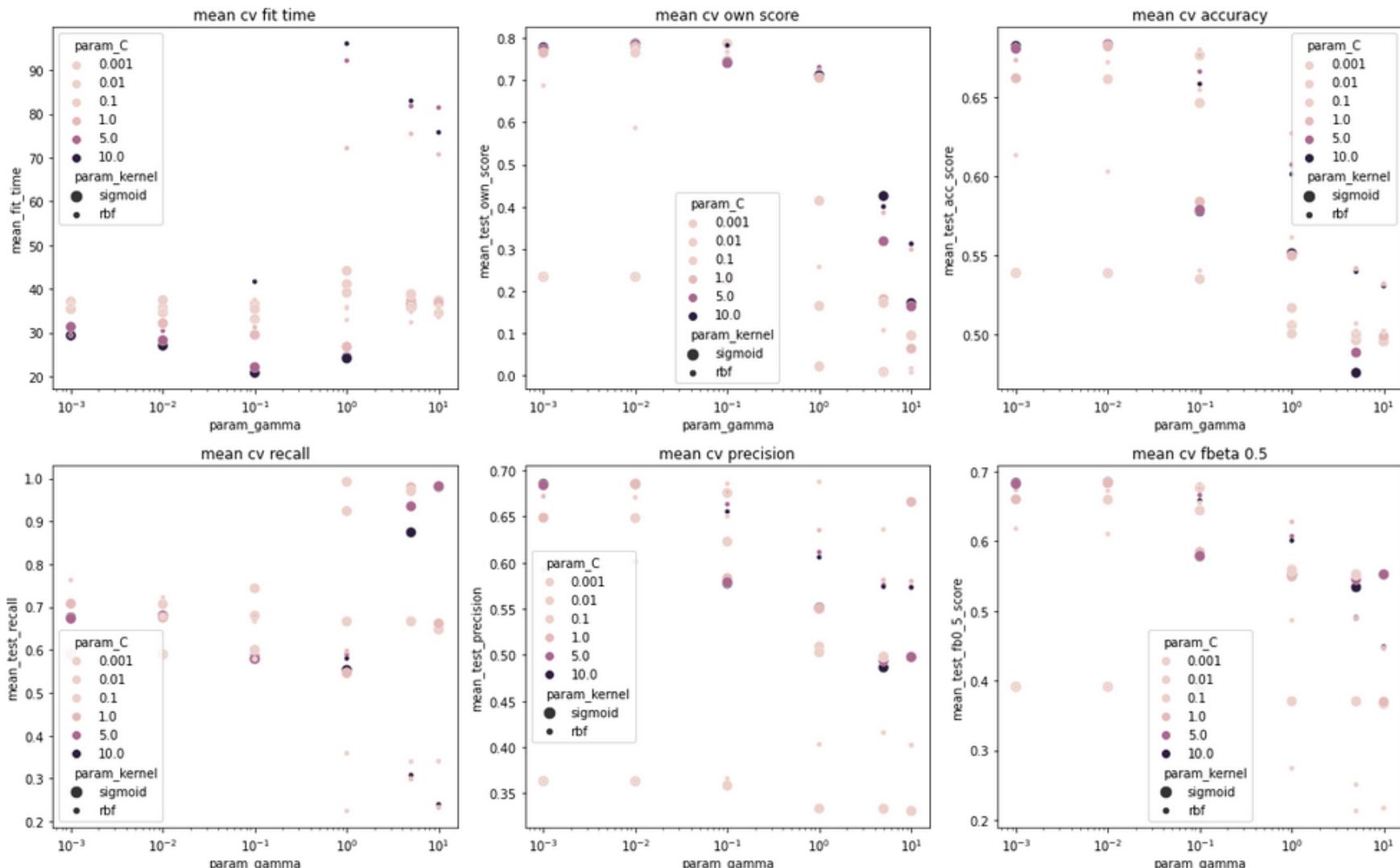
```
meilleur estimateur LinearSVC(C=1)
best score 0.79
best param {'C': 1, 'tol': 0.0001}
score prediction 0.79
roc auc: 0.68
```



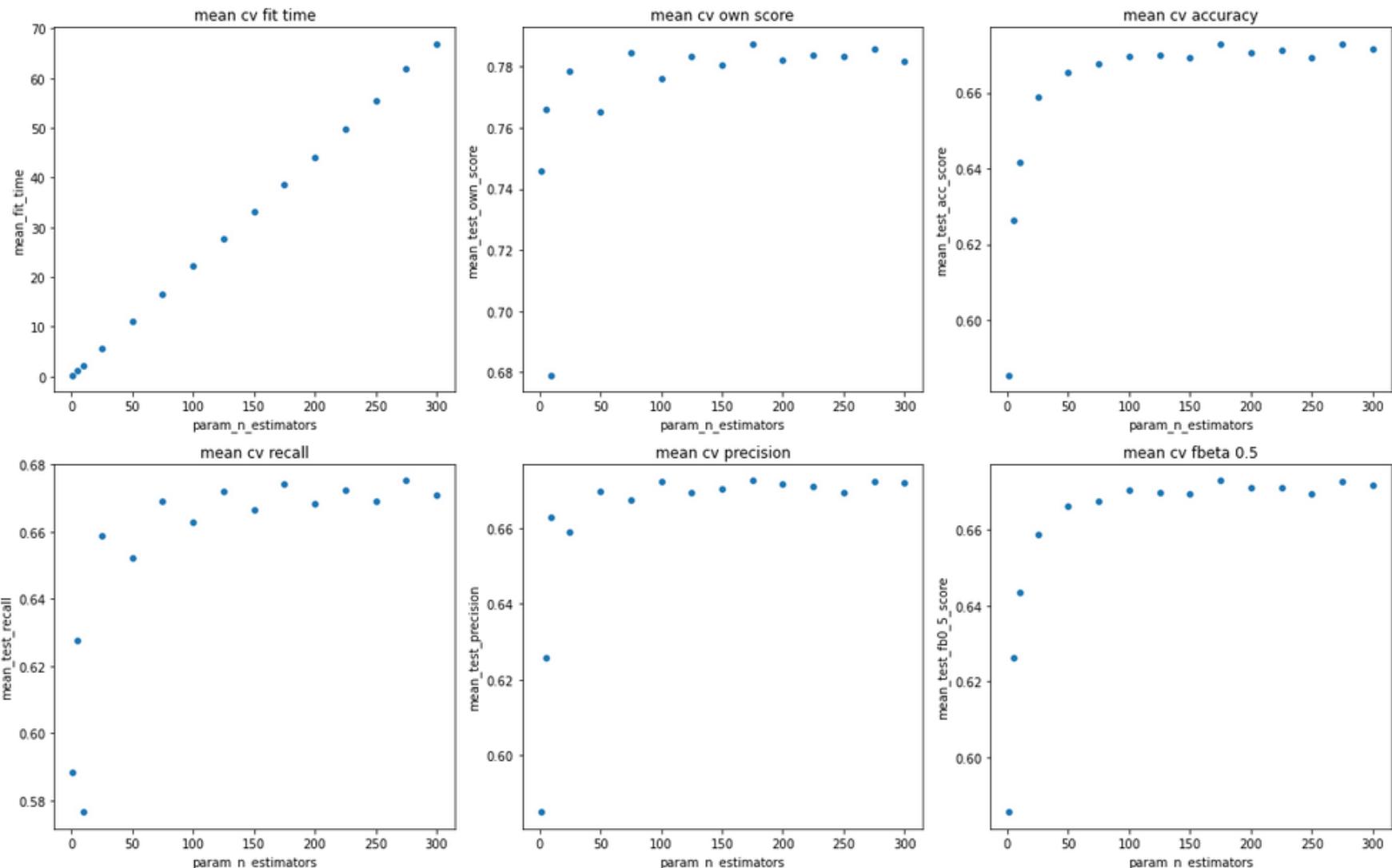
```

meilleur estimateur SVC(C=0.1, degree=2, gamma=0.1, kernel='sigmoid', tol=0.01)
best score 0.79
best param {'C': 0.1, 'gamma': 0.1, 'kernel': 'sigmoid'}
score prediction 0.76
roc auc: 0.63

```



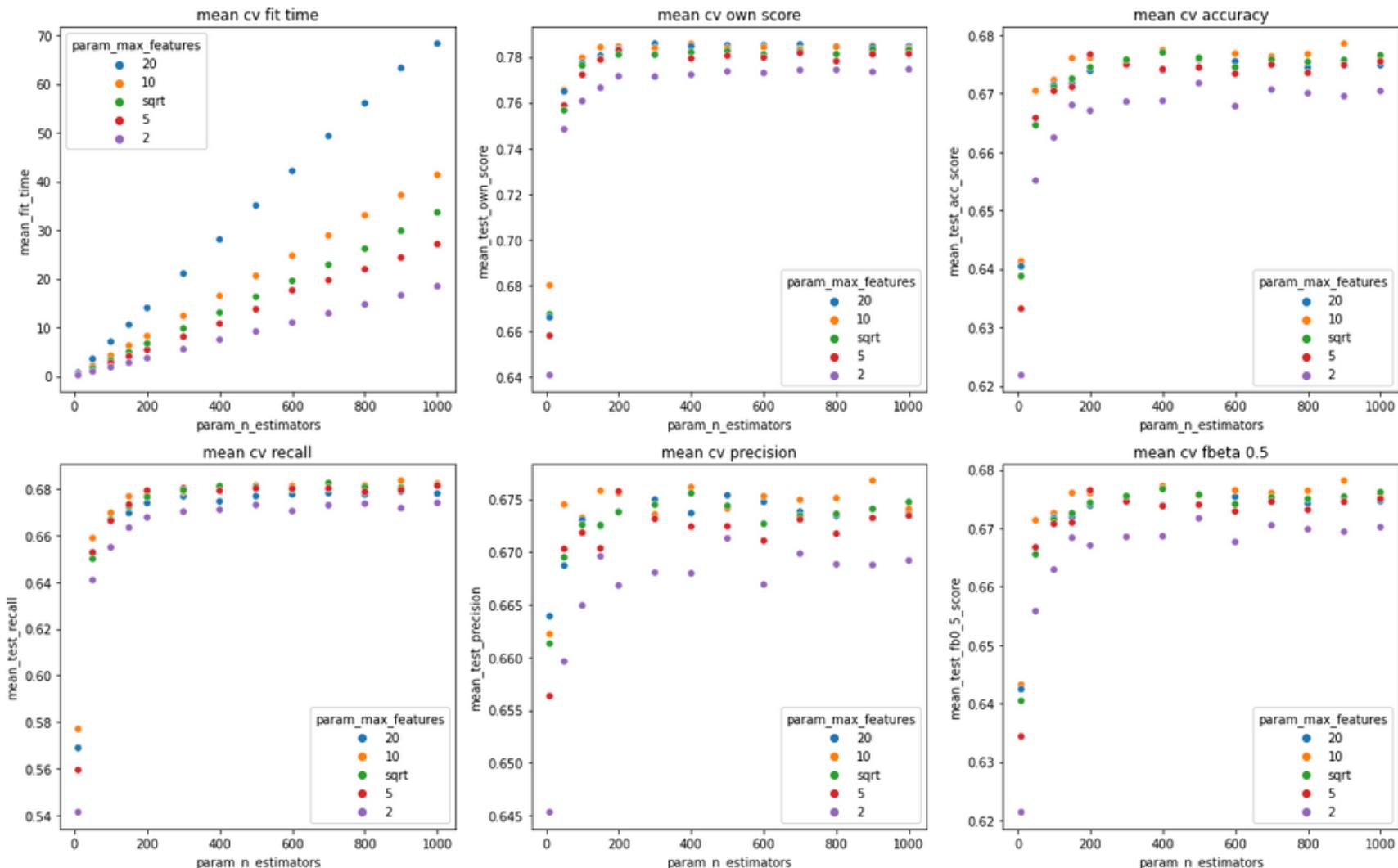
```
meilleur estimateur BaggingClassifier(n_estimators=175)
best score 0.79
best param {'n_estimators': 175}
score prediction 0.78
roc auc: 0.73
```



```

meilleur estimateur RandomForestClassifier(max_features=20, n_estimators=300)
best score 0.79
best param {'max_features': 20, 'n_estimators': 300}
score prediction 0.79
roc auc: 0.68

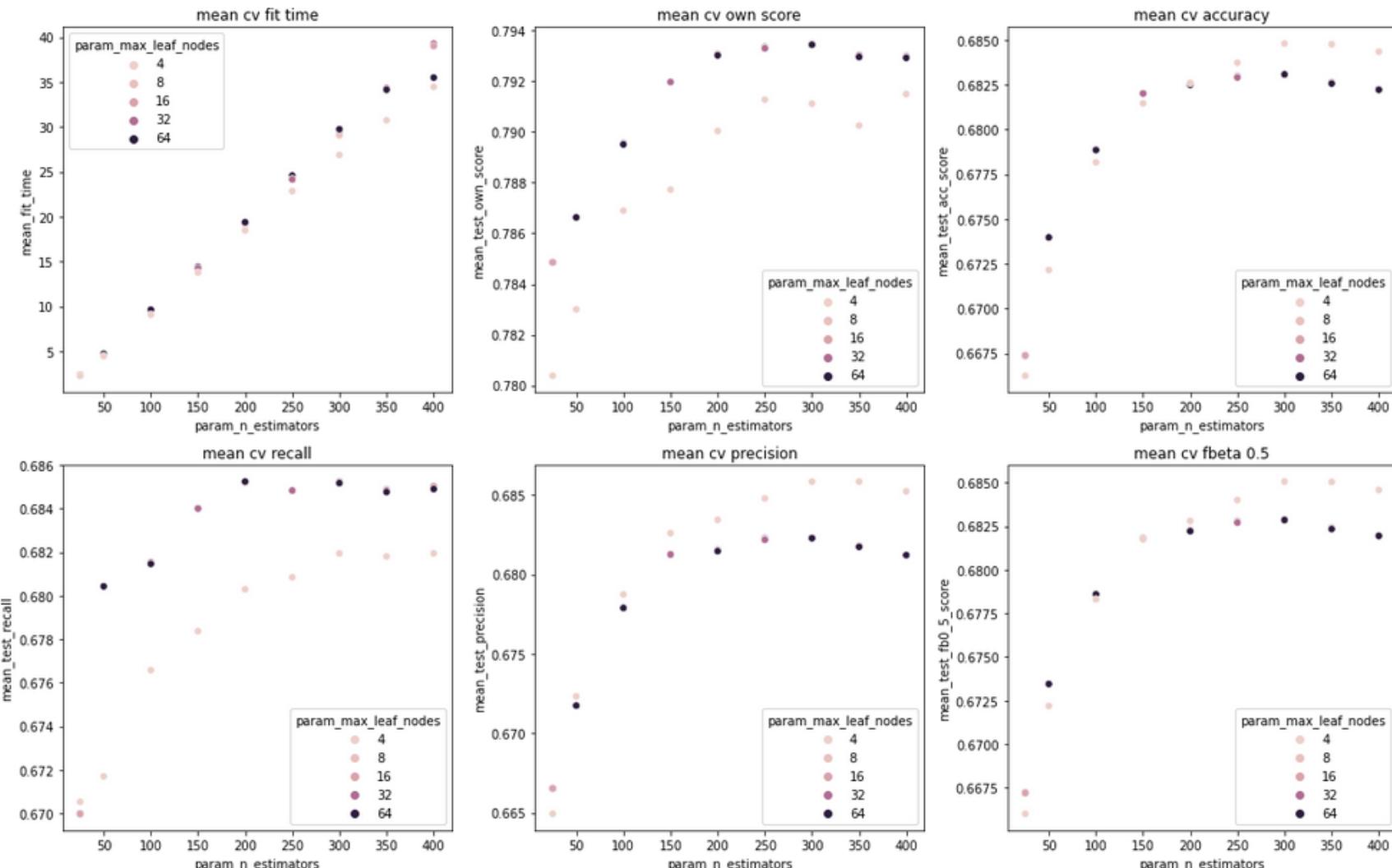
```



```

meilleur estimateur GradientBoostingClassifier(learning_rate=0.05, max_leaf_nodes=32,
                                              n_estimators=300, tol=0.001)
best score 0.79
best param {'learning_rate': 0.05, 'max_leaf_nodes': 32, 'n_estimators': 300, 'tol': 0.001}
score prediction 0.79
roc auc: 0.68

```



```

meilleur estimateur GradientBoostingClassifier(learning_rate=0.05, max_leaf_nodes=16,
                                              n_estimators=300, tol=0.001)
best score 0.80
best param {'learning_rate': 0.05, 'max_leaf_nodes': 16, 'n_estimators': 300, 'tol': 0.001}
score prediction 0.76

```

