OPENCLASSROOMS

Projet 04 :

Construisez un modèle de scoring

Mohamed A.

Contexte du projet :

Pour accorder un crédit à la consommation, l'entreprise souhaite mettre en œuvre un outil de "scoring crédit" qui calcule la probabilité qu'un client le rembourse ou non, puis classifie la demande : crédit accordé ou refusé. Elle souhaite donc développer un **algorithme de classification** pour aider à décider si un prêt peut être accordé à un client.

Les **chargés de relation client** seront les utilisateurs de l'outil de scoring. Puisqu'ils s'adressent aux clients, ils ont besoin que votre modèle soit **facilement interprétable**. Les chargés de relation souhaitent, en plus, disposer d'**une mesure de l'importance des variables** qui ont poussé le modèle à donner cette probabilité à un client.

Livrables

- Un **Jupyter Notebook** présentant les différentes parties de votre travail de modélisation.
  - Ce notebook doit pouvoir être utilisé par une autre personne, comme Michaël par exemple. Sa présentation et sa structuration doivent donc être soignées afin que le notebook puisse être pris en main par une personne autre que vous, sans que vous ayez à la former à son utilisation
- Une **présentation** (PowerPoint ou une alternative) :
  - Ce livrable vous servira à présenter votre approche méthodologique de modélisation de la problématique de scoring lors de la soutenance orale devant Michaël.

Contexte du projet :

1 – Presentation du dataset
2- Etapes du nettoyage et du traitement dataset
3- Modelisation par diverses approches
4- Identification des principales features par le module SHAPASH

1 – Presentation du dataset

Le dataset fourni est une table de Taille : 307 511 x 122 et qui contient des données relatives aux prets demandees par des particuliers avec un ensemble d'informations les concernant.

Le dataset contient des valeurs manquantes au niveau des colonnes ci-dessous :

```
# Features with missing values
[features for features in df.columns if df[features].i
```

```
['AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'NAME_TYPE_SUITE',
 'OWN_CAR_AGE',
 'OCCUPATION_TYPE',
 'CNT_FAM_MEMBERS',
 'EXT_SOURCE_1',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3',
 'OBS_30_CNT_SOCIAL_CIRCLE',
 'DEF_30_CNT_SOCIAL_CIRCLE',
 'OBS_60_CNT_SOCIAL_CIRCLE',
 'DEF_60_CNT_SOCIAL_CIRCLE',
 'DAYS_LAST_PHONE_CHANGE',
 'AMT_REQ_CREDIT_BUREAU_HOUR',
 'AMT_REQ_CREDIT_BUREAU_DAY',
 'AMT_REQ_CREDIT_BUREAU_WEEK',
 'AMT_REQ_CREDIT_BUREAU_MON',
 'AMT_REQ_CREDIT_BUREAU_QRT',
 'AMT_REQ_CREDIT_BUREAU_YEAR']
```

1 – Presentation du dataset

- Le tableau ci dessous regroupe l’ensemble des variables du dataset avec le nombre des valeurs manquantes , le type et la description de chaque variable.

- Les variables procurent diverses informations allant des montants du credit, revenus, annuités; ainsi que d’informations relatives au client telsque l’age, la sitation familiale, un apercu des biens , etc.…

- Ci apres le canevas Excel…..

1 – Etapes du nettoyage et du traitement dataset

col_ref	variable	count_of_valu	Missing_valu	Ratio_missing	dtyp	Nul_valu	Description
0	SK_ID_CURR	307511	0	0.0000	int64	0	ID of loan in our sample
1	TARGET	307511	0	0.0000	int64	282686	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
2	NAME_CONTRACT_TYPE	307511	0	0.0000	object	0	Identification if loan is cash or revolving
3	CODE_GENDER	307511	0	0.0000	object	0	Gender of the client
4	FLAG_OWN_CAR	307511	0	0.0000	object	0	Flag if the client owns a car
5	FLAG_OWN_REALTY	307511	0	0.0000	object	0	Flag if client owns a house or flat
6	CNT_CHILDREN	307511	0	0.0000	int64	215371	Number of children the client has
7	AMT_INCOME_TOTAL	307511	0	0.0000	float64	0	Income of the client
8	AMT_CREDIT	307511	0	0.0000	float64	0	Credit amount of the loan
9	AMT_ANNUITY	307499	12	0.0039	float64	0	Loan annuity
10	AMT_GOODS_PRICE	307233	278	0.0904	float64	0	For consumer loans it is the price of the goods for which the loan is given
11	NAME_TYPE_SUITE	306219	1292	0.4201	object	0	Who was accompanying client when he was applying for the loan
12	NAME_INCOME_TYPE	307511	0	0.0000	object	0	Clients income type (businessman, working, maternity leave,...)
13	NAME_EDUCATION_TYPE	307511	0	0.0000	object	0	Level of highest education the client achieved
14	NAME_FAMILY_STATUS	307511	0	0.0000	object	0	Family status of the client
15	NAME_HOUSING_TYPE	307511	0	0.0000	object	0	What is the housing situation of the client (renting, living with parents, ...)
16	REGION_POPULATION_RELATIVE	307511	0	0.0000	float64	0	Normalized population of region where client lives (higher number means the client lives in more populated region)
17	DAYS_BIRTH	307511	0	0.0000	int64	0	Client's age in days at the time of application
18	DAYS_EMPLOYED	307511	0	0.0000	int64	2	How many days before the application the person started current employment
19	DAYS_REGISTRATION	307511	0	0.0000	float64	80	How many days before the application did client change his registration
20	DAYS_ID_PUBLISH	307511	0	0.0000	int64	16	How many days before the application did client change the identity document with which he applied for the loan
21	OWN_CAR_AGE	104582	202929	65.9908	float64	2134	Age of client's car
22	FLAG_MOBIL	307511	0	0.0000	int64	1	Did client provide mobile phone (1=YES, 0=NO)
23	FLAG_EMP_PHONE	307511	0	0.0000	int64	55386	Did client provide work phone (1=YES, 0=NO)
24	FLAG_WORK_PHONE	307511	0	0.0000	int64	246203	Did client provide home phone (1=YES, 0=NO)
25	FLAG_CONT_MOBILE	307511	0	0.0000	int64	574	Was mobile phone reachable (1=YES, 0=NO)
26	FLAG_PHONE	307511	0	0.0000	int64	221080	Did client provide home phone (1=YES, 0=NO)
27	FLAG_EMAIL	307511	0	0.0000	int64	290069	Did client provide email (1=YES, 0=NO)
28	OCCUPATION_TYPE	211120	96391	31.3455	object	0	What kind of occupation does the client have
29	CNT_FAM_MEMBERS	307509	2	0.0007	float64	0	How many family members does client have
30	REGION_RATING_CLIENT	307511	0	0.0000	int64	0	Our rating of the region where client lives (1,2,3)
31	REGION_RATING_CLIENT_W_CITY	307511	0	0.0000	int64	0	Our rating of the region where client lives with taking city into account (1,2,3)
32	WEEKDAY_APPR_PROCESS_START	307511	0	0.0000	object	0	On which day of the week did the client apply for the loan
33	HOUR_APPR_PROCESS_START	307511	0	0.0000	int64	40	Approximately at what hour did the client apply for the loan
34	REG_REGION_NOT_LIVE_REGION	307511	0	0.0000	int64	302854	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)

1 – Etapes du nettoyage et du traitement dataset

col_ref	variable	count_of_valu	Missing_valu	Ratio_missing	dtyp	Nul_valu	Description
93	OBS_60_CNT_SOCIAL_CIRCLE	306490	1021	0.3320	float64	164666	How many observation of client's social surroundings with observable 60 DPD (days past due) default
94	DEF_60_CNT_SOCIAL_CIRCLE	306490	1021	0.3320	float64	280721	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
95	DAYS_LAST_PHONE_CHANGE	307510	1	0.0003	float64	37672	How many days before application did client change phone
96	FLAG_DOCUMENT_2	307511	0	0.0000	int64	307498	Did client provide document 2
97	FLAG_DOCUMENT_3	307511	0	0.0000	int64	89171	Did client provide document 3
98	FLAG_DOCUMENT_4	307511	0	0.0000	int64	307486	Did client provide document 4
99	FLAG_DOCUMENT_5	307511	0	0.0000	int64	302863	Did client provide document 5
100	FLAG_DOCUMENT_6	307511	0	0.0000	int64	280433	Did client provide document 6
101	FLAG_DOCUMENT_7	307511	0	0.0000	int64	307452	Did client provide document 7
102	FLAG_DOCUMENT_8	307511	0	0.0000	int64	282487	Did client provide document 8
103	FLAG_DOCUMENT_9	307511	0	0.0000	int64	306313	Did client provide document 9
104	FLAG_DOCUMENT_10	307511	0	0.0000	int64	307504	Did client provide document 10
105	FLAG_DOCUMENT_11	307511	0	0.0000	int64	306308	Did client provide document 11
106	FLAG_DOCUMENT_12	307511	0	0.0000	int64	307509	Did client provide document 12
107	FLAG_DOCUMENT_13	307511	0	0.0000	int64	306427	Did client provide document 13
108	FLAG_DOCUMENT_14	307511	0	0.0000	int64	306608	Did client provide document 14
109	FLAG_DOCUMENT_15	307511	0	0.0000	int64	307139	Did client provide document 15
110	FLAG_DOCUMENT_16	307511	0	0.0000	int64	304458	Did client provide document 16
111	FLAG_DOCUMENT_17	307511	0	0.0000	int64	307429	Did client provide document 17
112	FLAG_DOCUMENT_18	307511	0	0.0000	int64	305011	Did client provide document 18
113	FLAG_DOCUMENT_19	307511	0	0.0000	int64	307328	Did client provide document 19
114	FLAG_DOCUMENT_20	307511	0	0.0000	int64	307355	Did client provide document 20
115	FLAG_DOCUMENT_21	307511	0	0.0000	int64	307408	Did client provide document 21
116	AMT_REQ_CREDIT_BUREAU_HOUR	265992	41519	13.5016	float64	264366	Number of enquiries to Credit Bureau about the client one hour before application
117	AMT_REQ_CREDIT_BUREAU_DAY	265992	41519	13.5016	float64	264503	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
118	AMT_REQ_CREDIT_BUREAU_WEEK	265992	41519	13.5016	float64	257456	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
119	AMT_REQ_CREDIT_BUREAU_MON	265992	41519	13.5016	float64	222233	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
120	AMT_REQ_CREDIT_BUREAU_QRT	265992	41519	13.5016	float64	215417	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
121	AMT_REQ_CREDIT_BUREAU_YEAR	265992	41519	13.5016	float64	71801	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

1 – Etapes du nettoyage et du traitement dataset

TARGET COLUMN

['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'OWN_CAR_AGE', 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START', 'REG_REGION_NOT_LIVE_REGION', 'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE', 'EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE', 'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE', 'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE', 'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE', 'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRANCES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9', 'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT','AMT_REQ_CREDIT_BUREAU_YEAR']

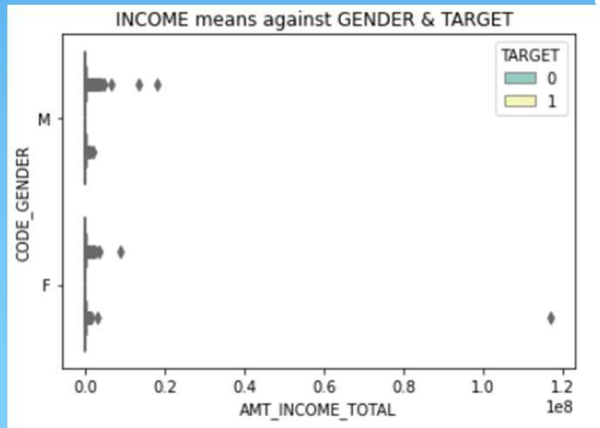
Check data types

1 – Etapes du nettoyage et du traitement dataset

	Action	Explication
1	Suppresion des colonnes allant de 'APARTMENTS_AVG' a 'EMERGENCYSTATE_MODE'	Repetition de 47 colonnes avec donnees differentes dans chaque colonnes
2	Definition colonne [AGE]	Transformation de la colonne [DAYS_BIRTH] exprimees en jours
3	Suppression des lignes avec gender XNA	4 valeurs dans tout le dataset
4	Remplacement de la valeur 117 000 000,00 pour la colonne [AMT_INCOME_TOTAL]	Voir dans le slide suivant
5	Remplacement des 'ANNUITY' missing values	Consideration du ratio moyen 'ratio_CRE_ANN' = 20

```
df['AGE_CLIENT'].unique()
```

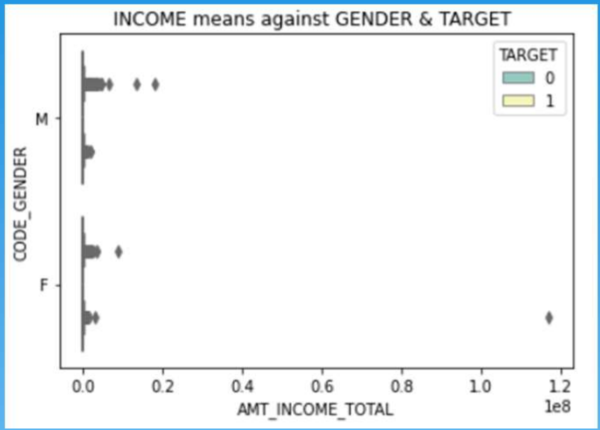
```
array([25, 45, 52, 54, 46, 37, 51, 55, 39, 27, 36, 38, 23, 35, 26, 48, 31,
       50, 40, 30, 68, 43, 28, 41, 32, 33, 47, 57, 65, 44, 64, 21, 59, 49,
       56, 62, 53, 42, 29, 67, 63, 61, 58, 60, 34, 22, 24, 66, 69, 20],
      dtype=int64)
```



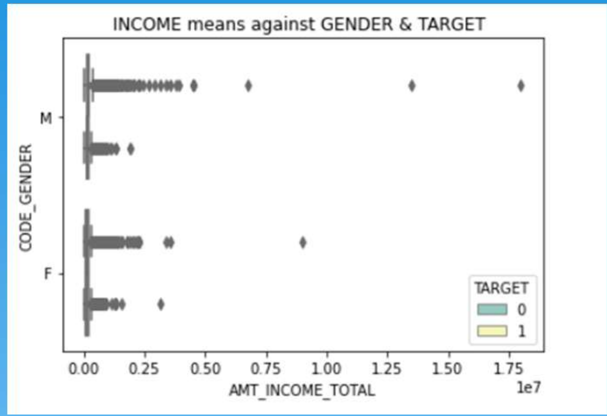

```
# Definition of some useful ratios
df['ratio_INC_ANN'] = df['AMT_INCOME_TOTAL']//df['AMT_ANNUITY'] # Defines payment annual capacity of annuity
df['ratio_CRE_ANN'] = df['AMT_CREDIT']//df['AMT_ANNUITY'] # Defines how many years needed to pay back the loan
df['ratio_CRE_INC'] = df['AMT_CREDIT']//df['AMT_INCOME_TOTAL'] # Defines credit value according to income capacity
```

1 – Traitement des outliers :

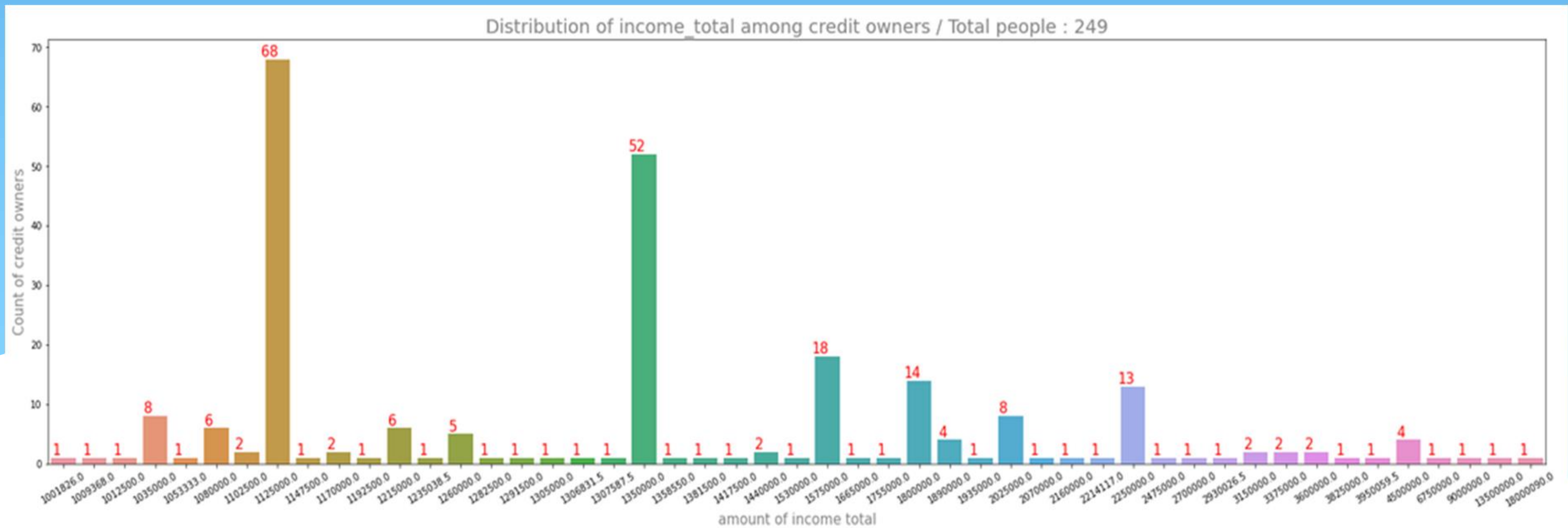
Raw data



After replacing main outlier value



Analyse de la distribution des loaners avec plus de 1 000 000 de INCOME



1 – Etapes du nettoyage et du traitement dataset

	Action	Explication
6	Remplacement missing values pour la colonne [NAME_TYPE_SUITE]	Remplacement par la variable 'Unnaccompagned' au vue de la distribution des valeurs existantes. Possibilite que les valeurs missing viennent du fait que le formulaire n'a pas ete rempli vu que le client etait seul
7	Remplacement missing values de la colonne ['OWN_CAR_AGE]	- Croisement avec la colonne[FLAG_OWN_CAR] et mise de la valeur (-1) pour celle conrespondant aux clients sans voiture - Remplcaement du reste des valeurs par la moyenne '4 annees'
8	Suppression colonne [OCCUPATION_TYPE]	- Impossibilite de repmlacement aleatoire et disponibilite d'info dans d'autres colonnes
9	Colonne [EXT_SOURCE_1] [EXT_SOURCE_2] [EXT_SOURCE_3]	- Pas de correlations trouvees avec les autres variables - Imputation d'une maniere aleatoire avec des valeurs de (0.05 a 0.9)
10	Remplacement missing value colonne [DAYS_LAST_PHONE_CHANGE]	- Repmlacement par la valeur (0) qui suppose que le client n'a pas change de numero de telephone
11	Remplacement missing values colonne [CNT_FAMILY_MEMBERS]	-Remplacement par des missing values (2 valeurs) par la valeur (1), vu que les deux personnes n'ont pas d'enfants et sont venus seules a la banque

1 – Etapes du nettoyage et du traitement dataset

	Action	Explication
12	Colonnes : OBS_30_CNT_SOCIAL_CIRCLE' & 'DEF_30_CNT_SOCIAL_CIRCLE'	- Remplacement des missing values par (0) ,ce qui implique qu'il n'y a pas de defaut ou observation a noter ou imputer. - Remplacement des valeurs superieurs a 31 par (30) , considerees comme outlier
13	Colonnes : OBS_60_CNT_SOCIAL_CIRCLE' & 'DEF_60_CNT_SOCIAL_CIRCLE	- Remplacement des missing values par (0) ,ce qui implique qu'il n'y a pas de defaut ou observation a noter ou imputer. - Remplacement des valeurs superieurs a 61 par (60) , considerees comme outlier
14	Colonne : [DAYS_EMPLOYED]	- Detection de la valeur outlier 365243 jours et remplacement par la valeur np.nan - Remplacement des valeurs missing par la moyenne relative a chaque tranche d'age
15	Colonne [FLAG_WORK_PHONE]	- colonne supprimee pour cause de repetition dans le dataset
16	Colonne [AMT_GOODS]	- Colonne supprimee vu qu'elle est fortement en correlation avec AMT_CREDIT
17	Colonne [CNT_CHILDREN]	- Colonne supprimee vu qu'elle est fortement en correlation avec CNT_FA:ILY_MEMBERS
18	Colonne [SK_ID]	- Colonne supprimee vu qu'elle n'est pas necessaire au modeling
19	Colonne [ORGNIZATION_TYPE]	- Colonne supprimee et on se contente de la classification fournie par la colonne [NAME_INCOME_TYPE], ce qui reduira le nombre de variable lors de l'encodage

1 – Etapes du nettoyage et du traitement dataset

	Action	Explication
20	Colonne [CNT_FAMILY_MEMBERS]	- Subdivision de la colonne en 4 categories de familes selon un critere de nombre : # FAM_UNI : up to 2 members # FAM_NOR : up to 4 members # FAM_NOM : up to 10 members # FAM_XXL : up to 20 members
21	Colonne [AMT_INCOME]	- Subdivision de la colonne en 5 categories sociales selon le montant de l'income : # POOR_CLASS : less than 35000 # AVG_CLASS : up to 150 000 # MED_CLASS : up to 500 000 # RICH_CLASS : up to 1 000 000 # JETSET_CLASS : up to 10 000 000
22	Colonne [AGE_CLIENT]	- - subdivision de la colonne en 5 categories selon l'age : - # YOUNG : up to 30 - # 30-TH : up to 40 - # 40-TH : up to 50 - # 50-TH : up to 60 - # SENIOR
23	Colonne [DAYS_LAST_PHONE_CHANGE]	Subdivision de la colonne en 5 categories selon la periode de changement : # Y1 : within first year # Y2 : within second year # Y3 : within third year # Y4 : within fourth year # +Y5 : after fifth year

1 – Etapes du nettoyage et du traitement dataset

	Action	Explication
24	Colonne [OWN_CAR_AGE]	- Merge des data de la colonne [FLAG_OWN_CAR] avec cette colonne et subdivision de la derniere en 6 categories selon l'age de la voiture : # NO_CAR : no car # NEW : less than 1 year # LIKE_NEW : up to 5 years # USED : up to 15 # OLD : up to 45 # COLLECTION : more than 45
25	Colonne [DAYS_EMPLOYED]	-subdivision de la colonne en 6 categories selon la date d'embauche : # NEWBEE : less than 1 year # PRE_PERMANENT : up to 2 years # PERMANENT : up to 5 years # EXPER_01 : up to 15 # EXPER_02 : up to 30 # PRE_RETIR : more than 30
26	Colonne [ratio_CRE_INC]	- Subdivision de la colonne en 5 categories selon la duree du pret : # EXTRA-SHORT : up to 7 years # SHORT : up to 20 years # MEDIUM : up to 30 years # LONG : up to 40 years # EXTRA-LONG : more than 40

1 – Etape Modelling : Mapping du travail realise durant le projet

Models used :

1- Logistic regression 2- Decision Tree 3- Light GBM

Preprocessing used :

* For numerical features :
1- MinMaxScaler
2- StandardScaler

*For categorical features :
1-OneHotEncode
2-Getdummies

Cross Validation : 3 folds used for all runs

Metrics : - AUC for all models
- Field score : user defined

Run models by using pipelines with considering set as umbalanced and balanced

GRID SEARCH CV : for Logistic Regression and LighGBM for hyperparameters tuning

Re-run models Logistic Regression and LighGBM with obtained best hyperparameters

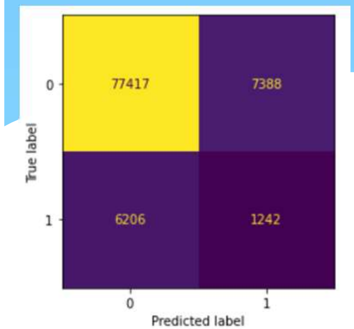
SHAP For global and local explainability

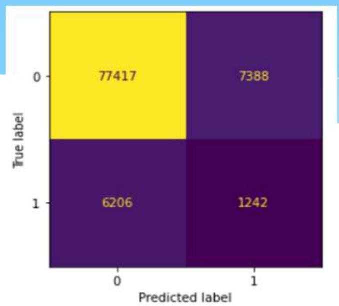
1 – Etape Modelling : Mapping du travail realise durant le projet

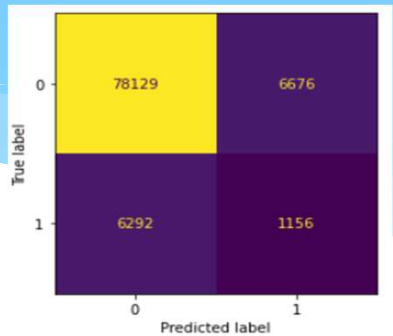
Results :

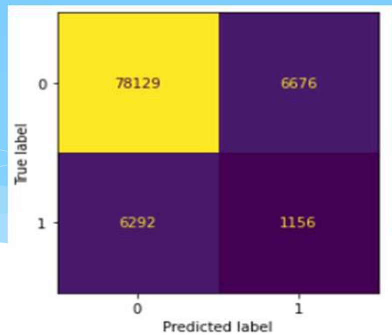
```
Model : DecisionTree
Umbalanced data set
-------------------------------------------------------
corss_val_field Scoring   : [5.96462264 5.91983455 5.89296777]
corss_val_mean : 5.925808320390186
-------------------------------------------------------
corss_val_accuracy Scoring   : [0.85165312 0.85199313 0.85291019]
corss_val_mean : 0.8521854804357561
-------------------------------------------------------
-------------------------------------------------------
Accuracy_score _train : 1.0
Accuracy_score _test  : 0.8526443584490477
-------------------------------------------------------
Precision score is    :  0.14391657010428738
Recall score is     : 0.16675617615467724
F1 score : 0.15449682796367706
-------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.5398193356452862
-------------------------------------------------------
roc_auc_score  _train is : 1.0
roc_auc_score  _test is : 0.5398193356452862
-------------------------------------------------------
Count of  "1" values in y_test    :  7448
Count of  "0" values in y_test    :  84805
-------------------------------------------------------
-------------------------------------------------------
```

```
Model : DecisionTree
Umbalanced data set adressed with "class_weight
-----------------------------------------------------------
corss_val_field Scoring    : [5.6974142   5.65246914 5.6645469 ]
corss_val_mean : 5.671476746820098
-----------------------------------------------------------
corss_val_accuracy Scoring    : [0.85658956 0.85900763 0.85860764]
corss_val_mean : 0.858068276425295
-----------------------------------------------------------
-----------------------------------------------------------
Accuracy_score _train : 1.0
Accuracy_score _test  : 0.8594300456353723
-----------------------------------------------------------
Precision score is    :  0.14759959141981613
Recall score is     : 0.1552094522019334
F1 score : 0.1513089005235602
-----------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.538243839360798
-----------------------------------------------------------
roc_auc_score  _train is : 1.0
roc_auc_score  _test is : 0.538243839360798
-----------------------------------------------------------
Count of  "1" values in y_test    :  7448
Count of  "0" values in y_test    :  84805
```







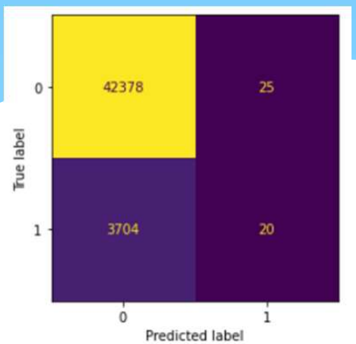


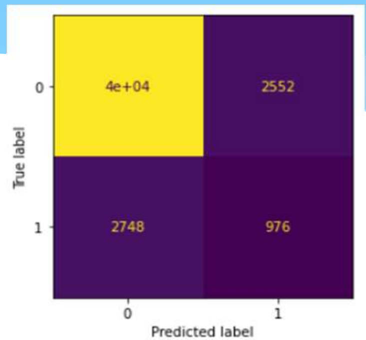
1 – Etape Modelling : Mapping du travail realise durant le projet

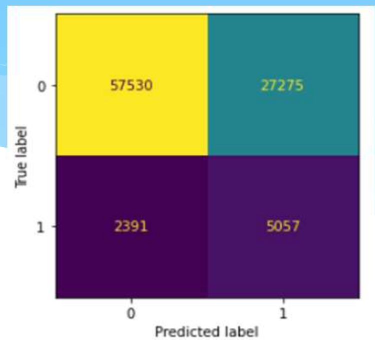
Results :

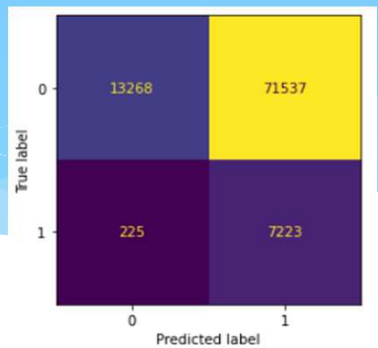
```
Model : Logistic regression
Umbalanced data set
------------------------------------------------------------
corss_val_field Scoring   : [1.0413693  1.0598115  1.04782031]
corss_val_mean : 1.0496670381077349
------------------------------------------------------
corss_val_accuracy Scoring   : [0.9193487  0.91926011 0.91921133]
corss_val_mean : 0.9192733821496102
----------------------------------------------------

----------------------------------------------------

Accuracy_score _train : 0.9193014002601576
Accuracy_score _test  : 0.919157976889891
----------------------------------------------------

Precision score is   :  0.4444444444444444
Recall score is    : 0.0053705692803437165
F1 score : 0.010612894667020428
----------------------------------------------------
-------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.6009496799835794
-------------------------------------------------------
roc_auc_score  _train is : 0.5026104047359637
roc_auc_score  _test is : 0.5023904941772328
-------------------------------------------------------
Count of  "1" values in y_test    :  3724
Count of  "0" values in y_test    :  42403
-------------------------------------------------------
```

```
Model : Logistic regression
Umbalanced data set considered with "class_weight"
--------------------------------------------------------------
corss_val_field Scoring   : [9.28701228 9.28010941 9.29752091]
corss_val_mean : 9.288214198354318
-----------------------------------------------------------
corss_val_accuracy Scoring   : [0.67895574 0.67898187 0.67927455]
corss_val_mean : 0.6790707207152451
-------------------------------------------------------

-------------------------------------------------------

Accuracy_score _train : 0.6799641353935351
Accuracy_score _test  : 0.6784278018059033
-------------------------------------------------------
Precision score is   :  0.15640851169120376
Recall score is    : 0.6789742212674543
F1 score : 0.25424836601307194
-------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.563121793561789
-------------------------------------------------------
roc_auc_score  _train is : 0.6837505896222662
roc_auc_score  _test is : 0.6786770168892545
Count of  "1" values in y_test    :  7448
Count of  "0" values in y_test    :  84805
--------------------------------------------------------------
```







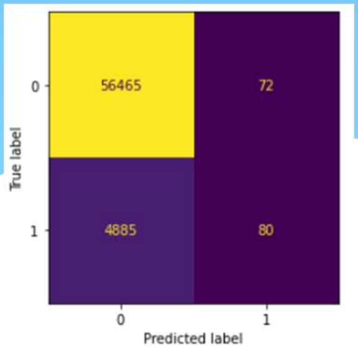


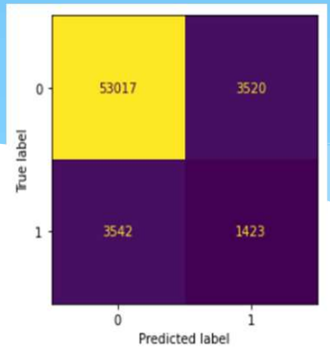
1 – Etape Modelling : Mapping du travail realise durant le projet

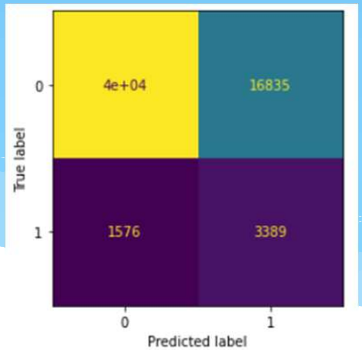
Results :

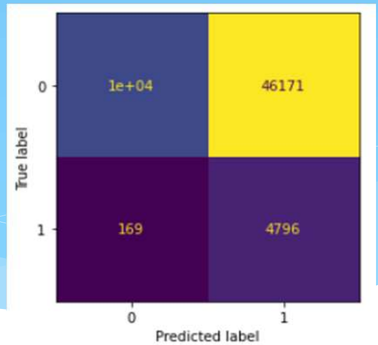
```
Model : Light GBM
Umbalanced data set
--------------------------------------------------------
corss_val_field Scoring   : [1.10356105 1.11451769 1.10767372]
corss_val_mean : 1.1085841517347583
--------------------------------------------------------
corss_val_accuracy Scoring   : [0.91945602 0.91949425 0.91926987]
corss_val_mean : 0.9194067125323867
--------------------------------------------------------
--------------------------------------------------------
Accuracy_score _train : 0.9201926790105892
Accuracy_score _test  : 0.9194009950895906
--------------------------------------------------------
------------------------------------------------------------
Precision score is   :  0.5263157894736842
Recall score is    : 0.016112789526686808
F1 score : 0.031268321282001174
------------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.6121730654297436
------------------------------------------------------------
roc_auc_score  _train is : 0.5094583428713457
roc_auc_score  _test is : 0.5074196436092319
------------------------------------------------------------
Count of  "1" values in y_test    :  4965
Count of  "0" values in y_test    :  56537
------------------------------------------------------------
```

```
Model : Light GBM
Umbalanced data set considered with "class_weight"
--------------------------------------------------------
corss_val_field Scoring   : [9.1939289  9.17018779 9.19912863]
corss_val_mean : 9.18774843940239
--------------------------------------------------------
corss_val_accuracy Scoring   : [0.70143313 0.70076681 0.70442528]
corss_val_mean : 0.7022084075330928
--------------------------------------------------------
--------------------------------------------------------
Accuracy_score _train : 0.7045263307656349
Accuracy_score _test  : 0.7006438814997886
--------------------------------------------------------
Precision score is   :  0.16757318037974683
Recall score is    : 0.6825780463242699
F1 score : 0.2690857120171503
--------------------------------------------------------
roc_auc_score  _with threshold: 0.8  is : 0.5746553447224296
--------------------------------------------------------
roc_auc_score  _train is : 0.7095581822319571
roc_auc_score  _test is : 0.6924042220584329
--------------------------------------------------------
Count of  "1" values in y_test    :  4965
Count of  "0" values in y_test    :  56537
--------------------------------------------------------
```









Grid search Results :

Logistic Regression	Best params: {'classifier__C': 0.001, 'classifier__penalty': 'l1', 'classifier__solver': 'liblinear'}

Light GBM	Best params: {‘num_leaves’ = 20, ‘max_depth’ = -1, ‘learning_rate’ = 0.1, ‘n_estimators’ =100'}

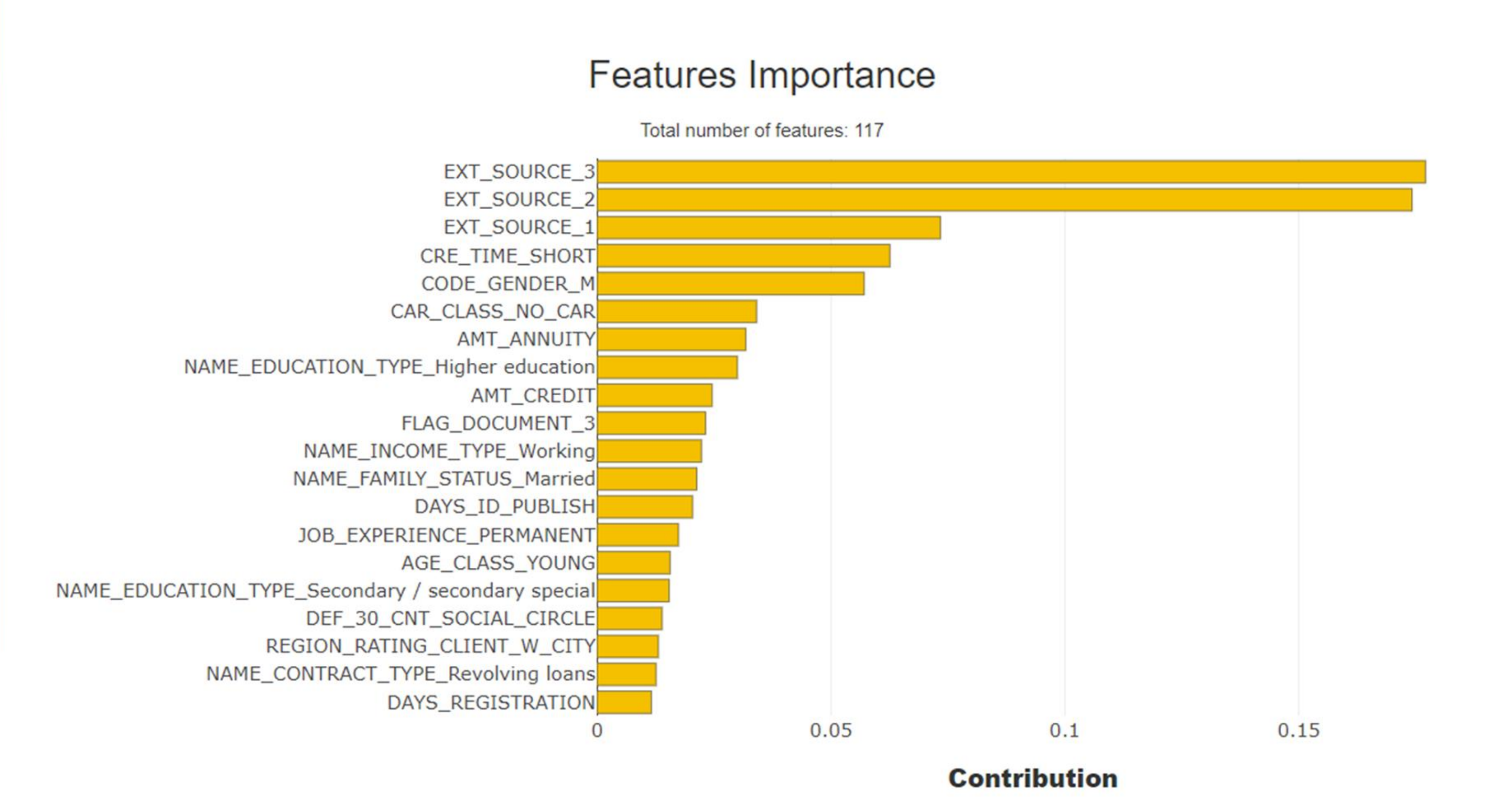
Light GBM (modele choisi)	Meilleur score AUC = 0.75

SHAPASH for model explainability :

1 – Global features explicability :

Le module SHPASH qui utilise les shapley values pour la determination de l’importance des features a ete utilise pour ce projet (modele :LightGBM)

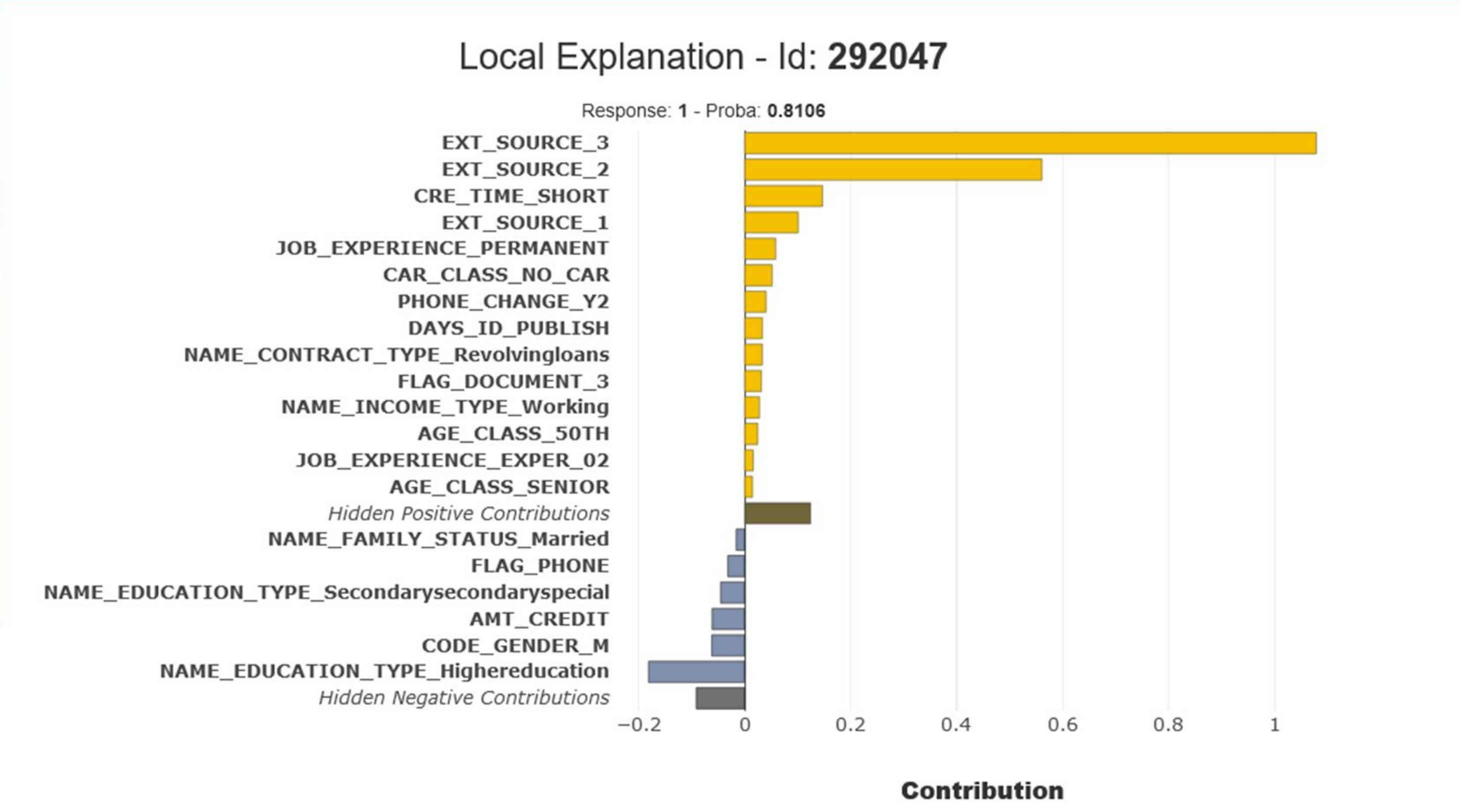
```
[ ]  xpl.plot.features_importance()
```



SHAPASH for model explainability :

2– local features explicability :

Le module SHPASH fournie la valeur des coefficient et de la contribution pour chaque enregistrement dans le dataset.
Ces donnees peuvent servir de base d’explication pour chaque decision d’attribution ou de refus d’un credit.



SHAPASH for model explainability :

– local features explicability :

Le module Shapash propose egalement une interface-web a utilisation facile qui permet de visualiser/selectionner/filtrer sur l’ensemble du dataset (voir ci dessous)