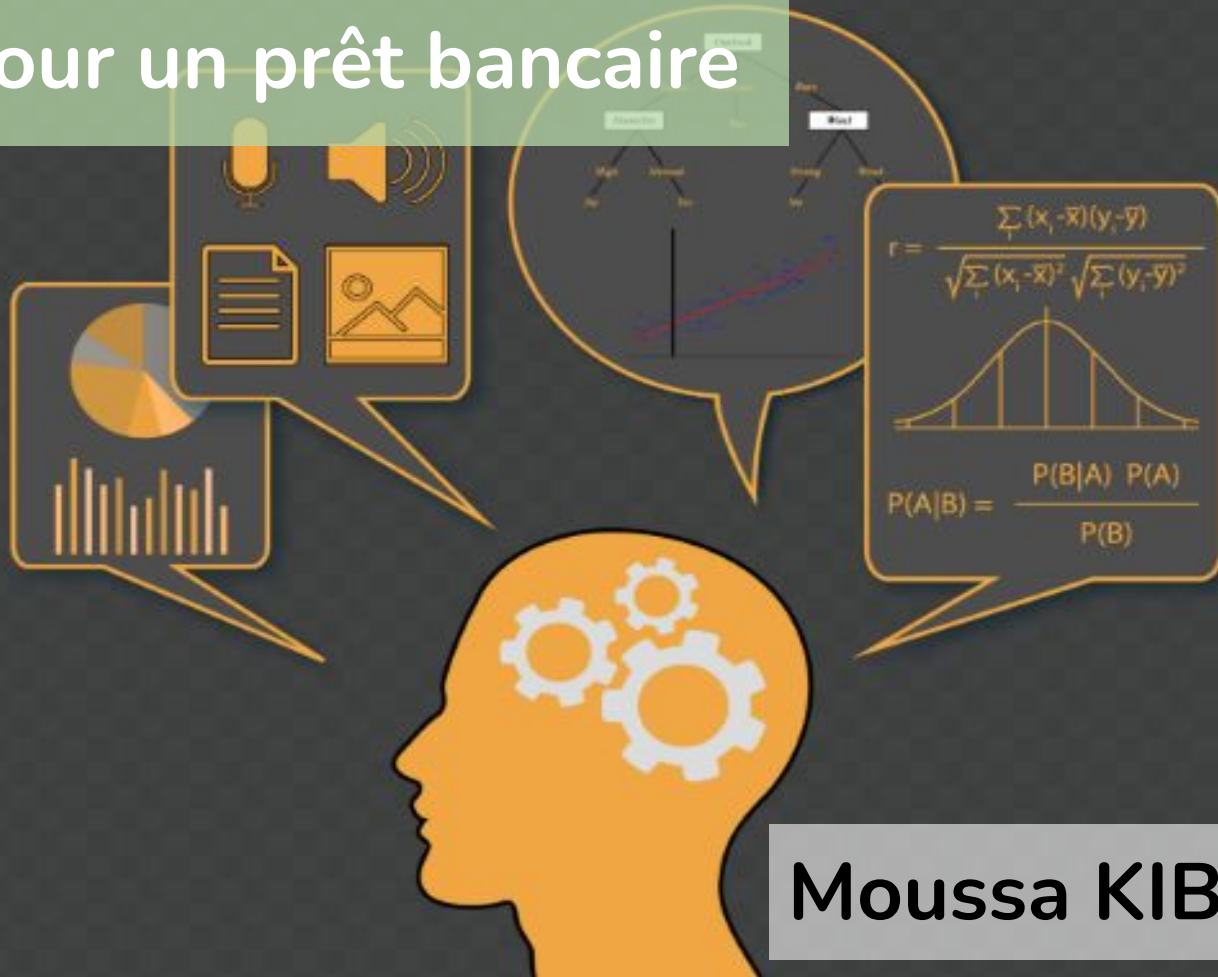


Construire un modèle de scoring pour un prêt bancaire



Moussa KIBALY



SOMMAIRE

1. CONTEXTE DU PROJET
2. PRÉSENTATION ET TRANSFORMATION DES DONNÉES
3. SYNTHÈSE DES MODÈLES D'APPRENTISSAGE
4. INTERPRÉTABILITÉ DU MODÈLE
5. CONCLUSION

1. CONTEXTE DU PROJET

Le société financière “Prêt à dépenser” souhaite développer un algorithme de scoring pour aider les chargés clientèles à attribuer ou non un crédit à un client. Pour cela, **une transformation des données pertinentes pour le modèle de scoring** sera faite sur le jeu de données “application_train.csv “ contenant des informations générales et financières sur le client et sur le crédit.



Les chargés de relation client vont utiliser le modèle de scoring. Ce dernier attribue **0** au client susceptible de rembourser son crédit et **1** à celui susceptible de ne pas le rembourser. De ce fait, le modèle sera optimisé en calculant sa performance sur la prédiction du score **1** car l'organisme de prêt cherche à éviter le défaut de paiement qui peut s'avérer coûteux en frais annexes.





L'amélioration de la performance du modèle sur la prédiction du score **1** est important pour le chargé clientèle. Pour ne pas perdre d'argent, il est essentiel que notre modèle soit fiable sur la prédiction des clients susceptibles de ne pas rembourser le crédit.



De plus, les chargés de clientèle souhaite connaître l'importance des variables (poids des variables) intervenants dans le modèle de scoring afin d'expliquer au client les raisons du refus du crédit, et ce de manière simple.



Grâce à l'analyse exploratoire et la transformation des données pertinentes, le conseiller bancaire sait déterminer de manière optimale le score d'un client. Il connaît de même les critères pertinents utilisés par le modèle.

La régression logistique optimisée permet d'obtenir une performance de 74.59% sur les prédictions des clients avec un score de 1.

2. PRÉSENTATION ET TRANSFORMATION DES DONNÉES



Le jeu de données des crédits consommation (fichier `application_train.csv`) contient 122 colonnes et 307511 lignes pour chaque demande de crédit. La donnée TARGET est à 0 quand le client a remboursé le crédit et 1 dans le cas contraire.

Le fichier contient des informations sur le score du client, le type de crédit, le genre du client, le montant de ses revenus, le montant du crédit, la date de naissance, la date d'embauche, des scores externes à la banque entre 0 et 1.

25.52% des cellules de notre jeu de données contiennent des données manquantes NaN qu'il faut traiter pour notre analyse. De plus, les variables catégorielles vont être transformées en numériques pour notre modèle de machine learning

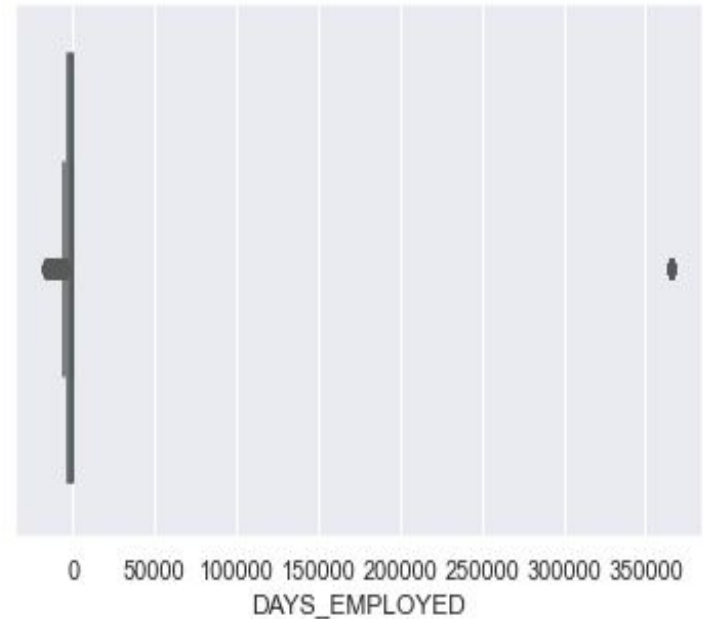
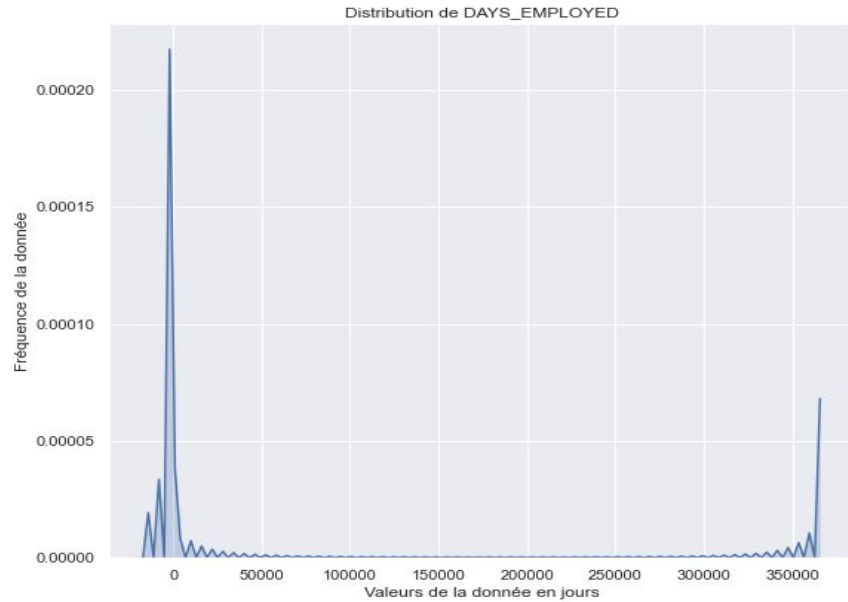
Indicateurs statistiques du jeu de données avant leur traitement

Out[81]:

CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
265992.000000	265992.000000	265992.000000	265992.000000	265992.000000
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
0.007000	0.034362	0.267395	0.265474	1.899974
0.110757	0.204685	0.916002	0.794056	1.869295
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	0.000000	0.000000	3.000000
9.000000	8.000000	27.000000	261.000000	25.000000

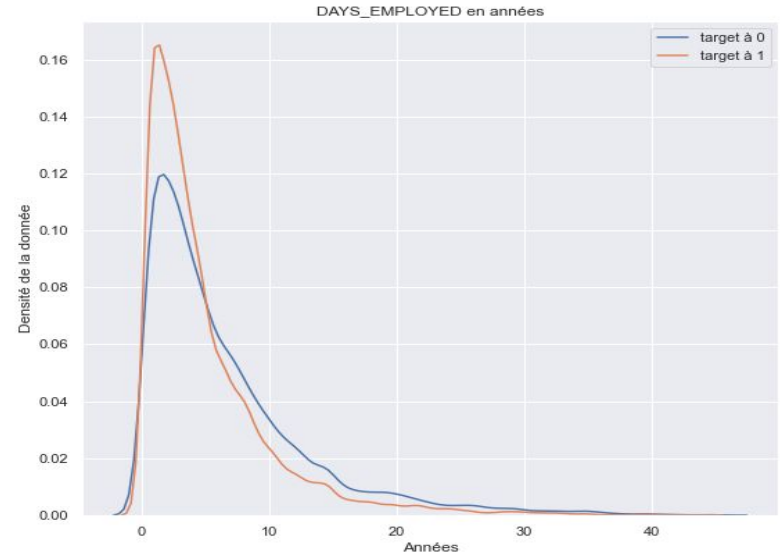
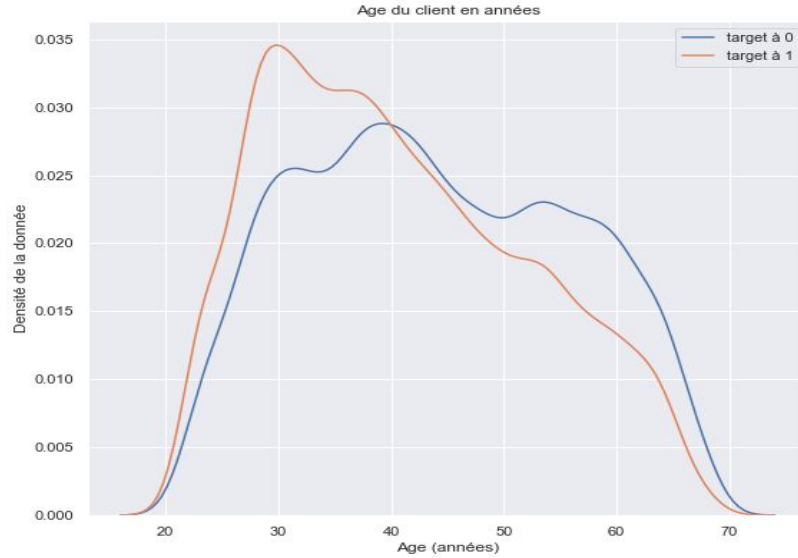
```
count    49650.000000
mean     53960.558348
std      132136.153466
min      -16069.000000
25%      -2479.000000
50%      -1119.000000
75%       -337.000000
max       365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```

On observe des valeurs aberrantes pour la donnée **DAYS_EMPLOYED** avec un outlier à 365243 jours ce qui correspond à 1000 ans.



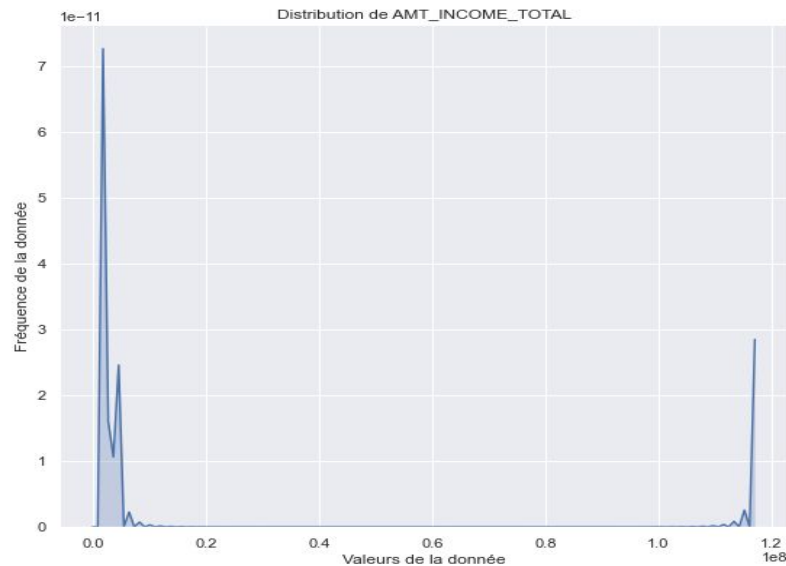
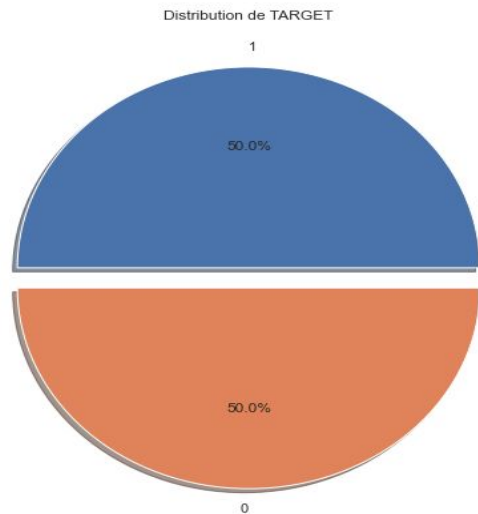
Ces graphiques montrent bien l'outlier de la variable DAYS_EMPLOYED. On peut facilement voir que sa valeur est supérieure à $1.5 \times \text{écart interquartile (Q3 - Q1)}$. Cette variable va être traitée en remplaçant la valeur outlier par NaN puis en substituant les NaN par la médiane de la colonne DAYS_EMPLOYED.

Distribution des variables quantitatives



On note bien l'influence de la date de naissance DAYS_BIRTH sur le target. Les client plus jeunes ont tendance à ne pas rembourser leur crédit. Des clients avec de faibles périodes d'activité professionnelle ont une tendance à ne pas rembourser.

8.1% des clients ne remboursent pas leur crédit (TARGET=1) dans le jeu initial. Cela va engendrer un sur-apprentissage du modèle sur le TARGET=0 car il est prédominant. C'est pour cela qu'on a équilibré les scores 0 et 1. La distribution des revenus des clients indique une majorité de clients avec des revenus annuels < 200 000.



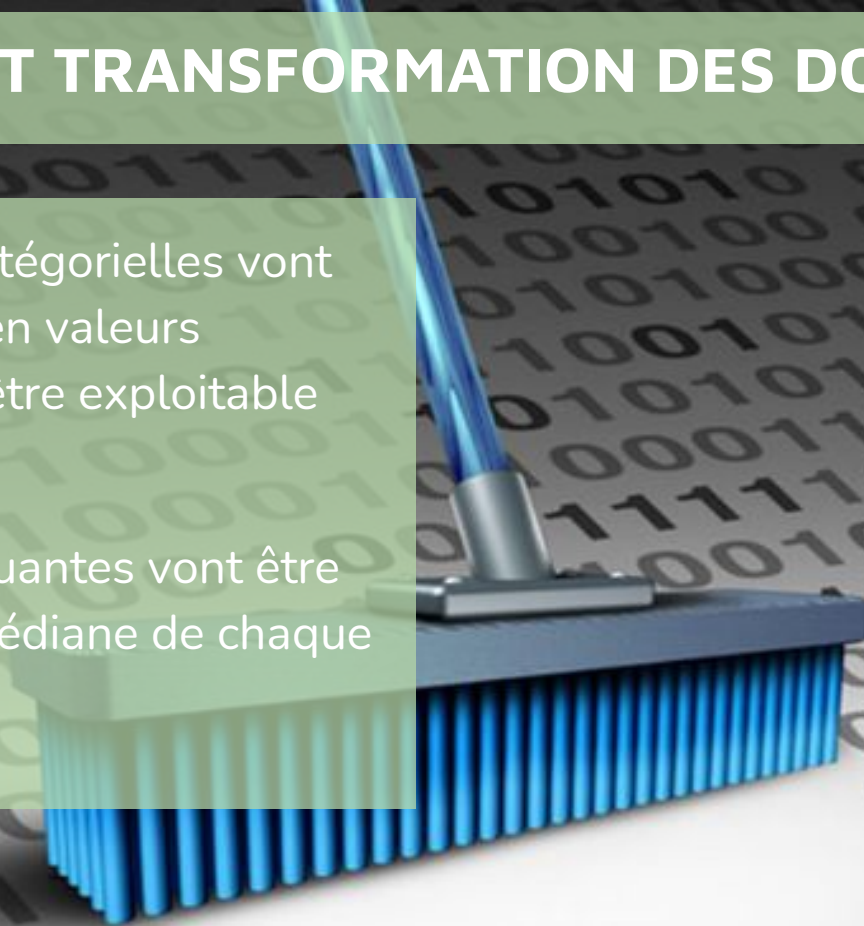
On observe une asymétrie vers la droite des distributions du montant des revenus annuels avec des outliers supérieurs à 80 000 000. Cela est mis en évidence par l'indicateur élevé du **skewness empirique** qui est positif et égale à **391.56 pour fr.** Le **kurtosis empirique** de cette variable est positif et élevé ce qui signifie un aplatissement moins important que la distribution normale.

De même pour le montant du crédit, la distribution est asymétrique vers la droite mais de manière moins important que le revenu avec en majorité un montant du crédit inférieur à 1 000 000. **Le skewness est positif mais seulement de 1.23.**

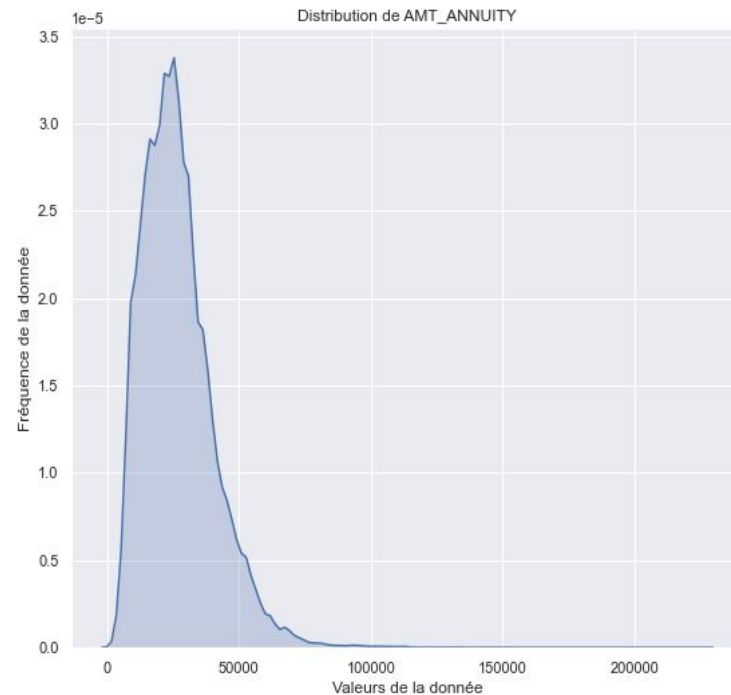
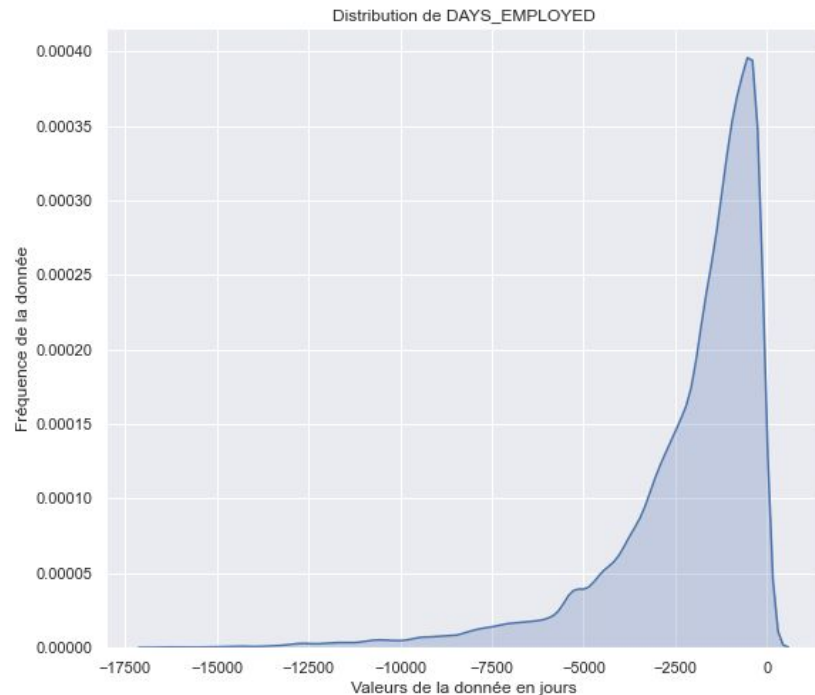
NETTOYAGE ET TRANSFORMATION DES DONNÉES

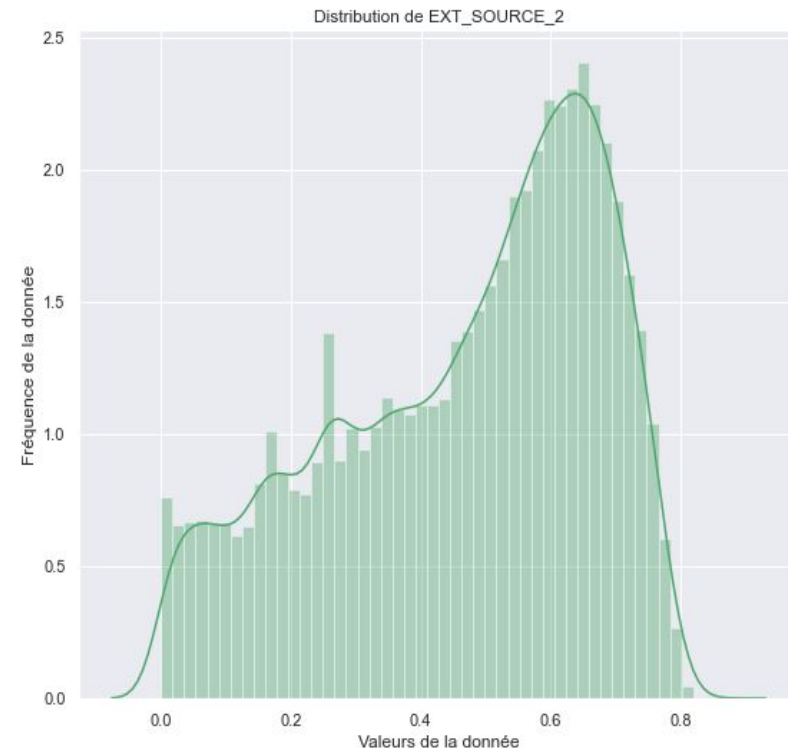
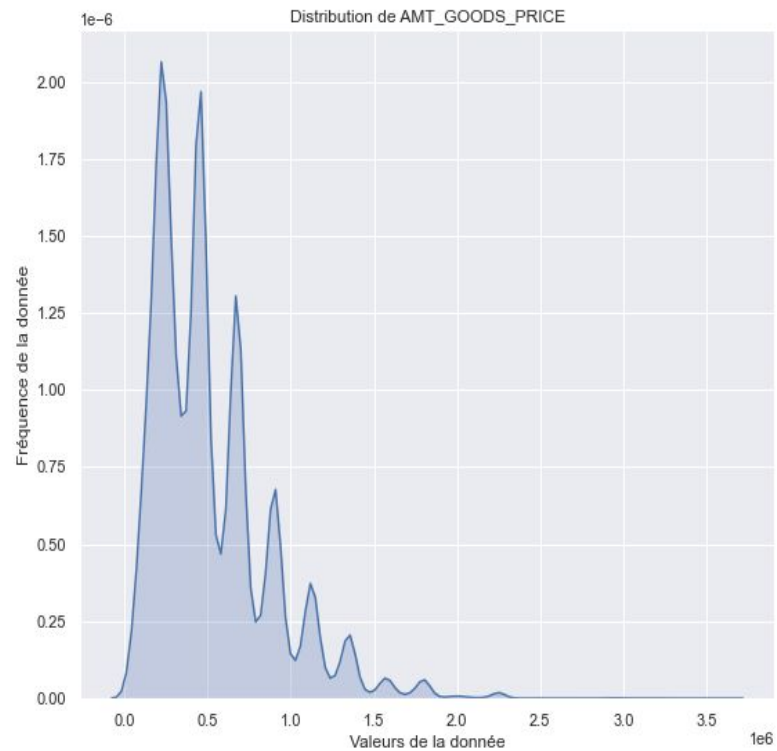
Les 16 variables catégorielles vont être transformées en valeurs numériques afin d'être exploitable par le modèle.

Les données manquantes vont être substitués par la médiane de chaque variable.

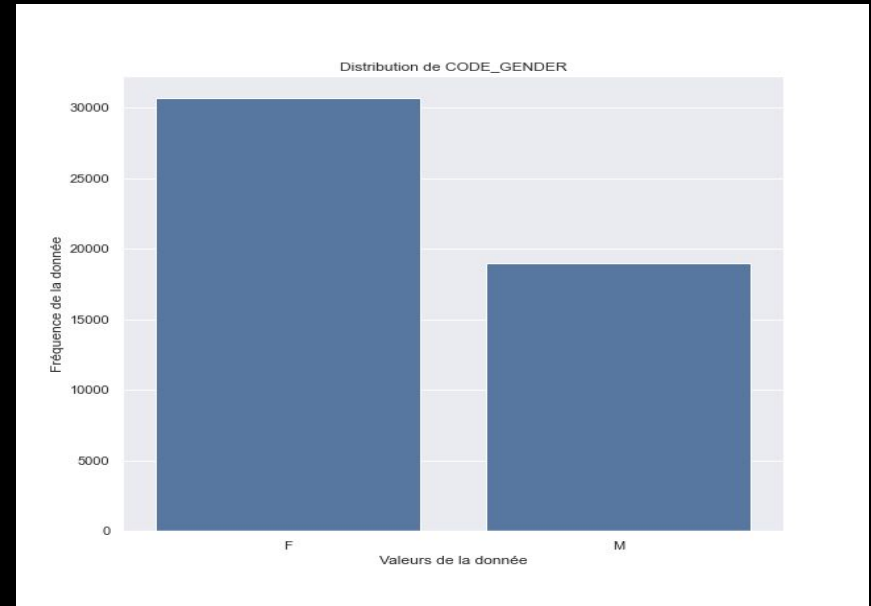
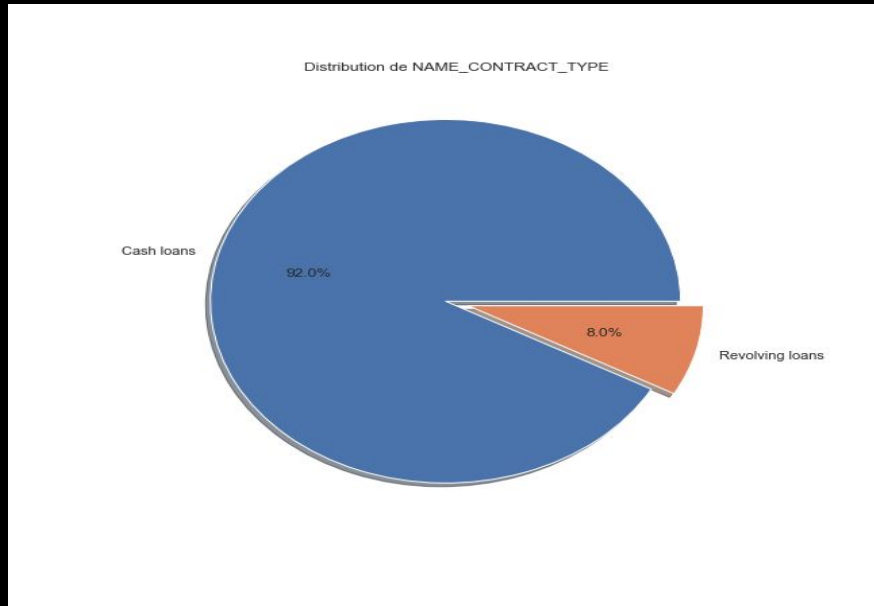


ANALYSE STATISTIQUE UNIVARIÉE (APRÈS NETTOYAGE)

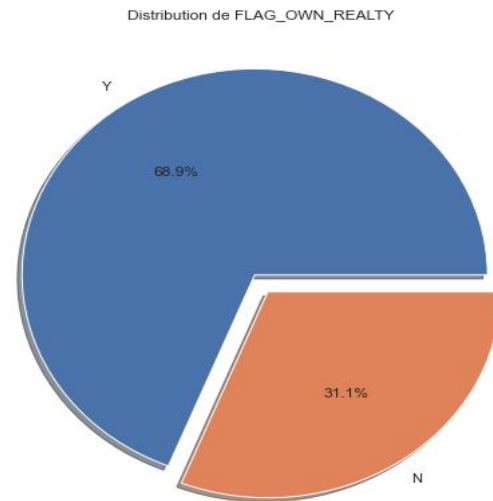
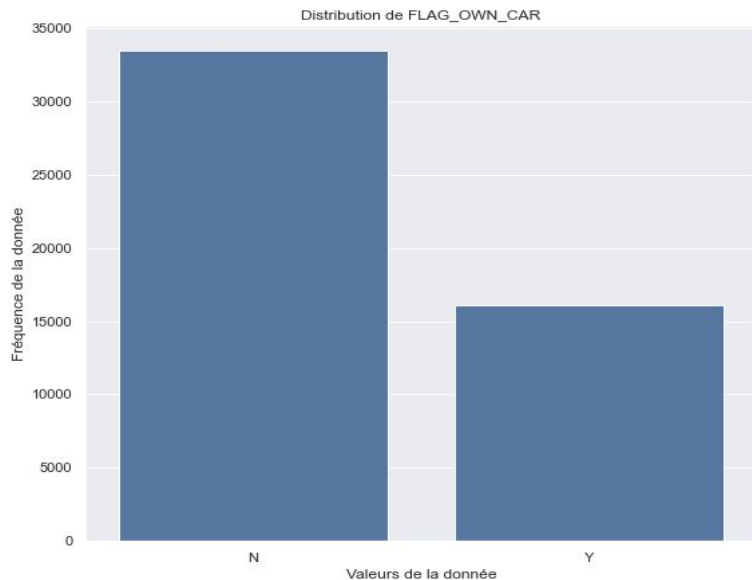





Distribution des variables qualitatives




La grande majorité des crédits consommation sont en cash et le reste en revolving. Au niveau du genre des clients, les femmes sont en plus grand nombre pour les demandes de crédit consommation. Une majorité de client ne possède pas de voiture.



Une majorité de client ne possède pas de voiture mais possède un bien immobilier.



```
NAME_CONTRACT_TYPE      2
CODE_GENDER              3
FLAG_OWN_CAR             2
FLAG_OWN_REALTY          2
NAME_TYPE_SUITE          7
NAME_INCOME_TYPE         8
NAME_EDUCATION_TYPE      5
NAME_FAMILY_STATUS       6
NAME_HOUSING_TYPE        6
OCCUPATION_TYPE          18
WEEKDAY_APPR_PROCESS_START 7
ORGANIZATION_TYPE        58
FONDKAPREMONT_MODE       4
HOUSETYPE_MODE           3
WALLSMATERIAL_MODE       7
EMERGENCYSTATE_MODE      2
dtype: int64
```

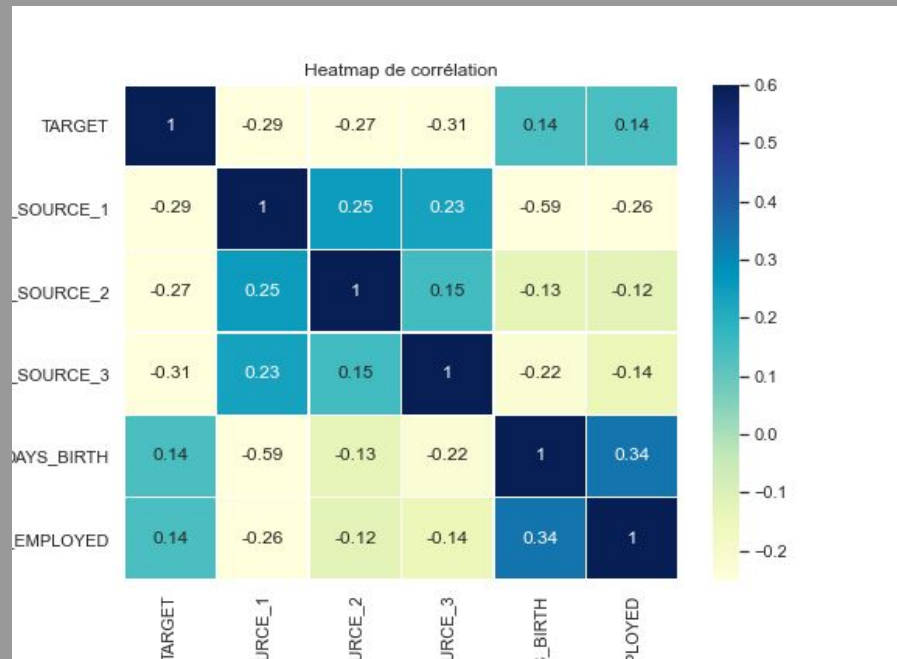


Il est important de convertir les données des 16 colonnes catégorielles en données numériques afin d'être exploitable par le modèle de machine learning. Les techniques utilisées sont le **label encoding** pour les variables avec moins de 2 catégories et le **one-hot encoding** pour les variables avec plus de 2 catégories.

Corrélations entre le TARGET et les autres variables numériques. Le coefficient de corrélation de Pearson est affiché dans l'ordre croissant.

```
Fortes corrélations positives:
REG_CITY_NOT_WORK_CITY      0.090432
FLAG_DOCUMENT_3             0.091306
DAYS_ID_PUBLISH             0.091394
FLAG_EMP_PHONE              0.093918
NAME_EDUCATION_TYPE_Secondary / secondary special 0.096019
CODE_GENDER_M               0.099756
REGION_RATING_CLIENT        0.104343
DAYS_LAST_PHONE_CHANGE      0.106097
REGION_RATING_CLIENT_W_CITY 0.106968
NAME_INCOME_TYPE_working    0.107601
DAYS_EMPLOYED               0.145531
DAYS_BIRTH                  0.146185
TARGET                      1.000000
FLAG_DOCUMENT_10            NaN
FLAG_DOCUMENT_12            NaN
Name: TARGET, dtype: float64
```

```
Fortes corrélations négatives:
EXT_SOURCE_3                -0.308691
EXT_SOURCE_1                -0.282392
EXT_SOURCE_2                -0.273252
NAME_EDUCATION_TYPE_Higher education -0.110105
CODE_GENDER_F               -0.099714
NAME_INCOME_TYPE_Pensioner  -0.094465
ORGANIZATION_TYPE_XNA       -0.093928
DAYS_EMPLOYED_ANOM          -0.093928
FLOORSMAX_AVG               -0.086028
FLOORSMAX_MEDI              -0.086023
FLOORSMAX_MODE              -0.085846
AMT_GOODS_PRICE             -0.076221
EMERGENCYSTATE_MODE_No      -0.075442
HOUSETYPE_MODE_block of flats -0.075145
REGION_POPULATION_RELATIVE  -0.067971
```



Corrélations POSITIVES entre le TARGET et DAYS_BIRTH, DAYS_EMPLOYED.

Corrélations NÉGATIVES entre EXT_SOURCE_3,2,1 et TARGET.

$$= 2x^3 + x^2 - 12x + 9$$

$$= (x - 1)(2x^2 + 3x - 9)$$

Pour augmenter la corrélation des variables EXT_SOURCE et DAYS_BIRTH avec TARGET, on va former de nouvelles variables qui sont des polynômes des variables initiales.

Dans notre cas, on va utiliser le degré $d=3$ (ex : EXT_SOURCE_1 \times EXT_SOURCE_2²).



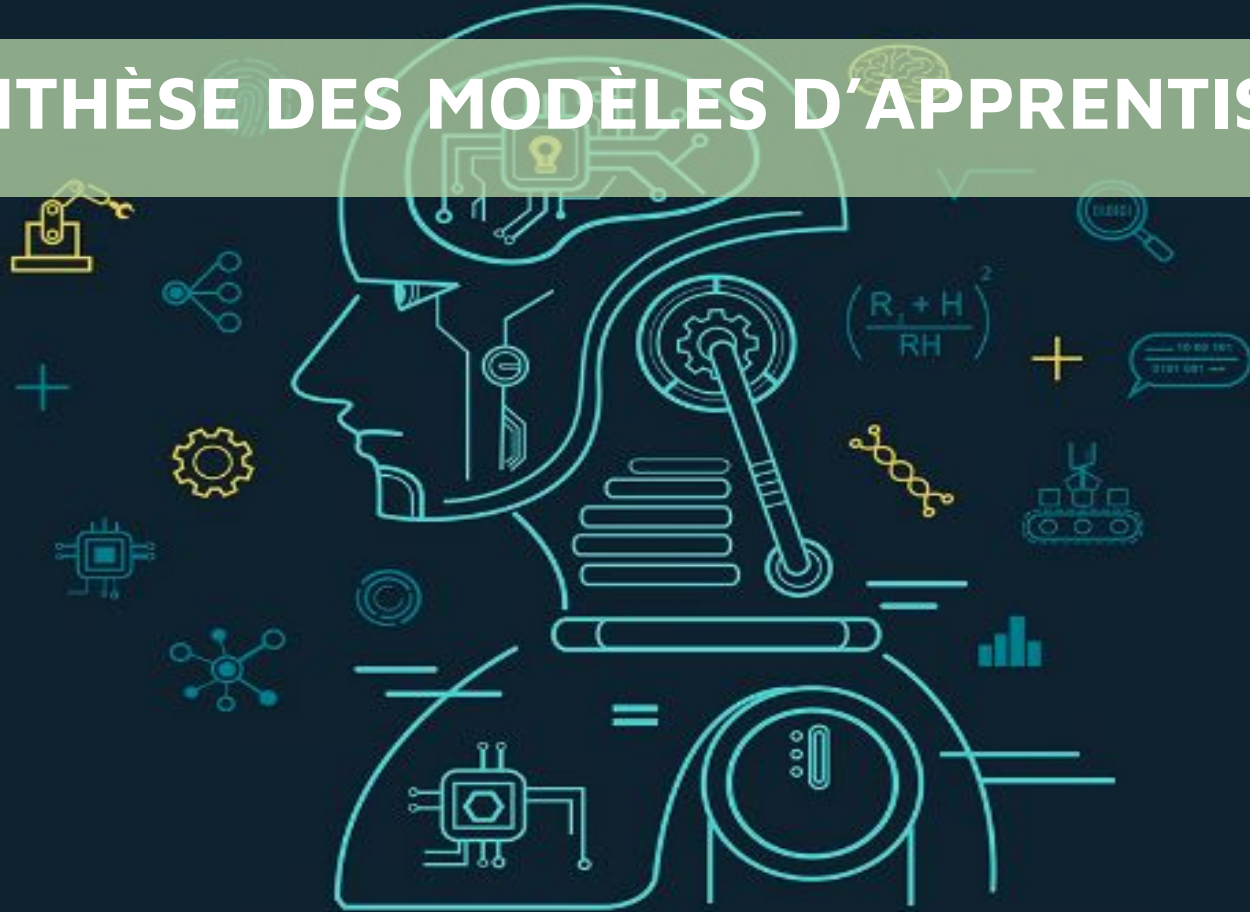
Corrélations entre le TARGET et les 35 variables polynomiales.

```
'EXT_SOURCE_1',  
'EXT_SOURCE_2',  
'EXT_SOURCE_3',  
'DAYS_BIRTH',  
'EXT_SOURCE_1^2',  
'EXT_SOURCE_1 EXT_SOURCE_2',  
'EXT_SOURCE_1 EXT_SOURCE_3',  
'EXT_SOURCE_1 DAYS_BIRTH',  
'EXT_SOURCE_2^2',  
'EXT_SOURCE_2 EXT_SOURCE_3',  
'EXT_SOURCE_2 DAYS_BIRTH',  
'EXT_SOURCE_3^2',  
'EXT_SOURCE_3 DAYS_BIRTH',  
'DAYS_BIRTH^2',  
'EXT_SOURCE_1^3',  
'EXT_SOURCE_1^2 EXT_SOURCE_2',  
'EXT_SOURCE_1^2 EXT_SOURCE_3',  
'EXT_SOURCE_1^2 DAYS_BIRTH',  
'EXT_SOURCE_1 EXT_SOURCE_2^2',  
'EXT_SOURCE_1 EXT_SOURCE_2 EXT_SOURCE_3',  
'EXT_SOURCE_1 EXT_SOURCE_2 DAYS_BIRTH',  
'EXT_SOURCE_1 EXT_SOURCE_3^2',  
'EXT_SOURCE_1 EXT_SOURCE_3 DAYS_BIRTH',  
'EXT_SOURCE_1 DAYS_BIRTH^2',  
'EXT_SOURCE_2^3',  
'EXT_SOURCE_2^2 EXT_SOURCE_3',  
'EXT_SOURCE_2^2 DAYS_BIRTH',  
'EXT_SOURCE_2 EXT_SOURCE_3^2',  
'EXT_SOURCE_2 EXT_SOURCE_3 DAYS_BIRTH',  
'EXT_SOURCE_2 DAYS_BIRTH^2',  
'EXT_SOURCE_3^3',  
'EXT_SOURCE_3^2 DAYS_BIRTH',  
'EXT_SOURCE_3 DAYS_BIRTH^2',  
'DAYS_BIRTH^3']
```

```
EXT_SOURCE_1 EXT_SOURCE_2 EXT_SOURCE_3    -0.358344  
EXT_SOURCE_2 EXT_SOURCE_3                -0.353985  
EXT_SOURCE_2^2 EXT_SOURCE_3              -0.336984  
EXT_SOURCE_2 EXT_SOURCE_3^2              -0.329260  
EXT_SOURCE_1 EXT_SOURCE_3                -0.303534  
EXT_SOURCE_1 EXT_SOURCE_2                -0.301382  
EXT_SOURCE_1 EXT_SOURCE_2^2              -0.295058  
EXT_SOURCE_1 EXT_SOURCE_3^2              -0.289979  
EXT_SOURCE_2                             -0.272979  
EXT_SOURCE_3                             -0.272946  
Name: TARGET, dtype: float64  
EXT_SOURCE_1 EXT_SOURCE_3 DAYS_BIRTH      0.291258  
EXT_SOURCE_1 EXT_SOURCE_2 DAYS_BIRTH      0.293417  
EXT_SOURCE_2 EXT_SOURCE_3 DAYS_BIRTH      0.345236  
TARGET                                     1.000000  
1                                           NaN  
Name: TARGET, dtype: float64
```

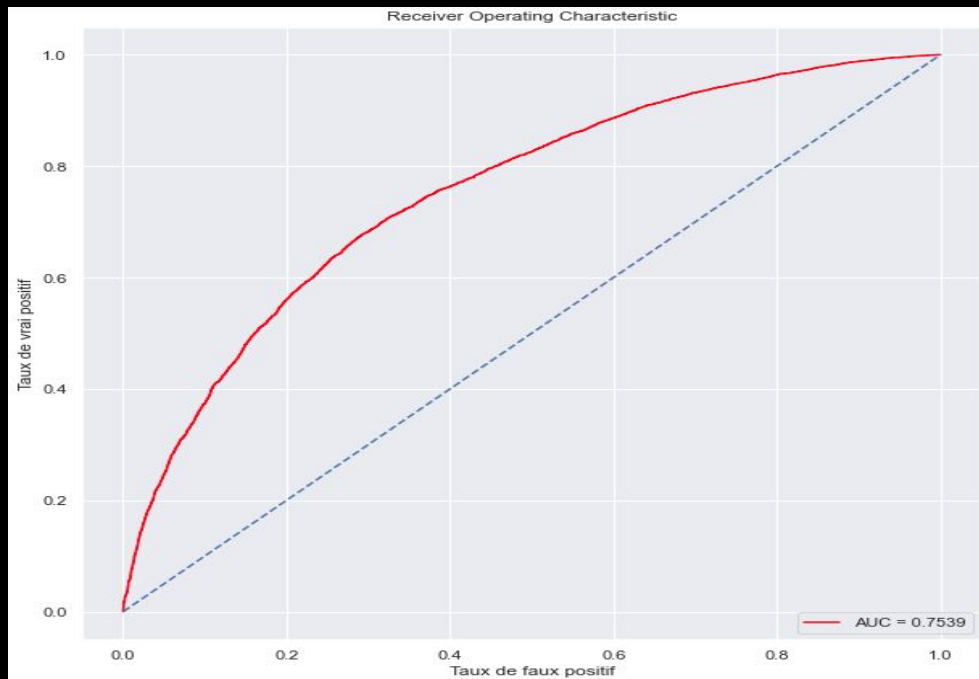
Corrélations NÉGATIVES augmentées entre **EXT_SOURCE_2 * EXT_SOURCE_3** et **TARGET**. Corrélations POSITIVES augmentées entre **EXT_SOURCE_2 * EXT_SOURCE_3 * DAYS-BIRTH** et **TARGET**. Ces nouvelles variables peuvent améliorer la performance du modèle de scoring.

3. SYNTHÈSE DES MODÈLES D'APPRENTISSAGE



MACHINE LEARNING

On va entraîner notre premier modèle linéaire de classification binaire : **LA RÉGRESSION LOGISTIQUE**



F-mesure : 0.6887
F50-mesure : 0.6970
Aire sous la courbe ROC: 0.7539

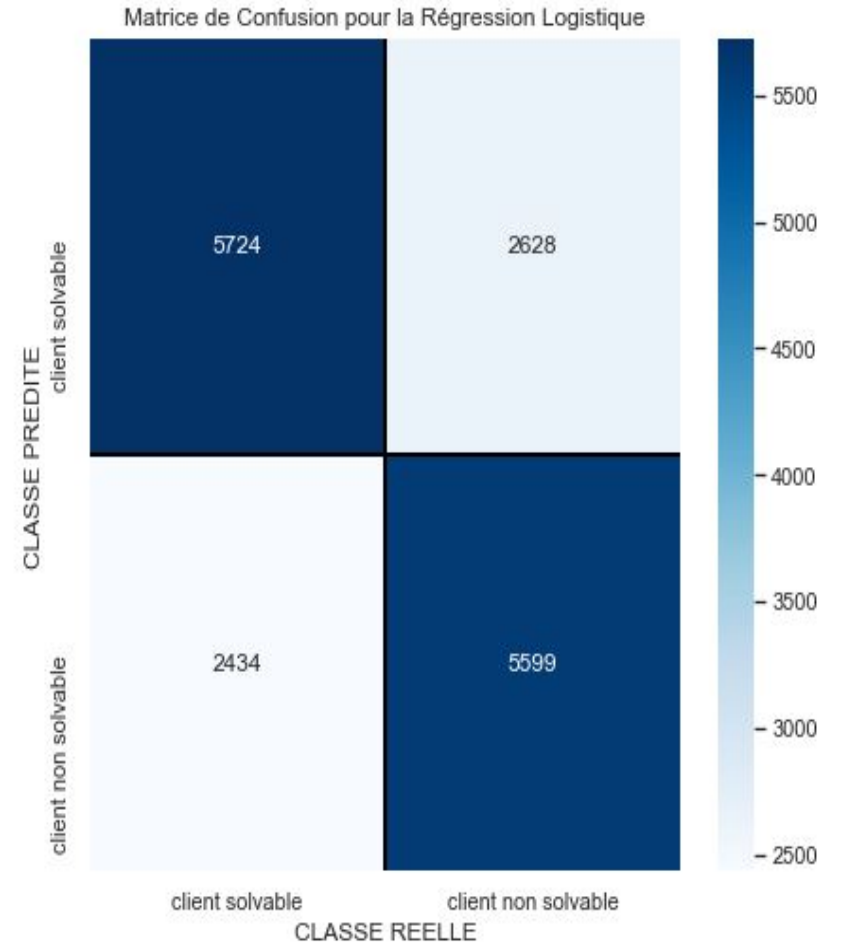
On entraîne le modèle après avoir normalisé entre 0 et 1 les variables du fichier `X_train`. On mesure la performance du modèle en calculant l'aire sous la courbe ROC (Receiver Operator Characteristic).

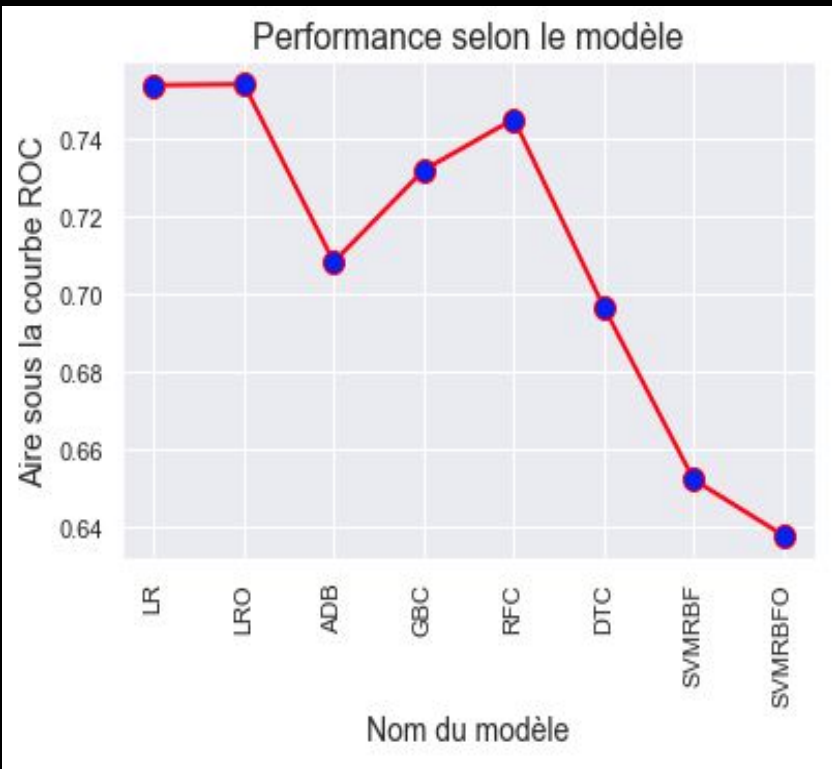
On obtient une performance de **0.7536** qui sera notre référence. Le modèle optimisé porte cette valeur à **0.7539**.

Optimized Logistic Regression Classifier report:

	precision	recall	f1-score	support
0	0.70	0.69	0.69	8352
1	0.68	0.70	0.69	8033
accuracy			0.69	16385
macro avg	0.69	0.69	0.69	16385
weighted avg	0.69	0.69	0.69	16385

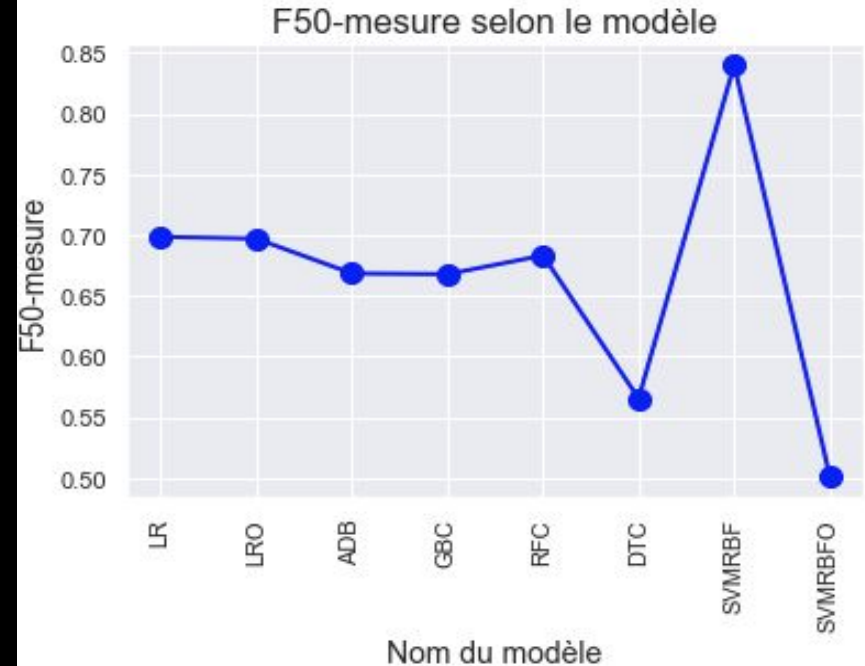
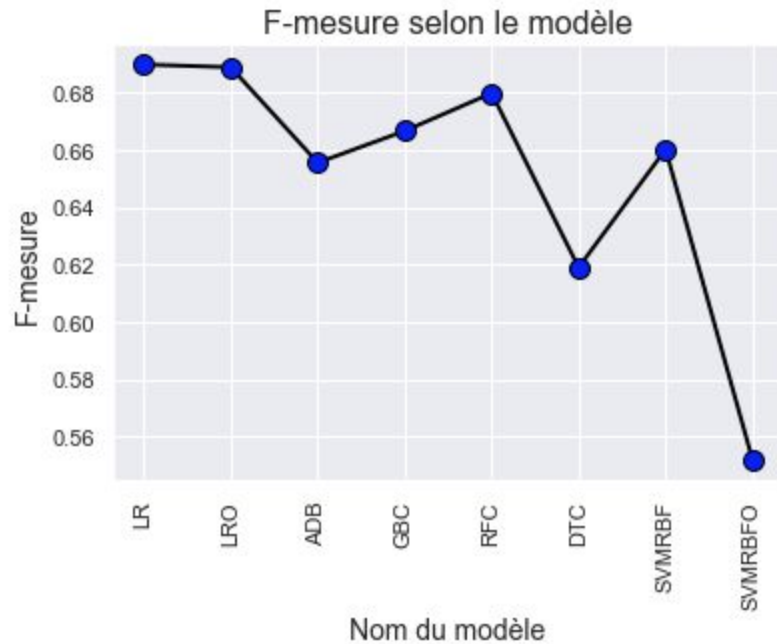
On utilise des métriques telles que la F-mesure, F50-mesure, le rapport de classification ci-dessus et le heatmap à droite afin d'évaluer notre modèle à maximiser les vrais positifs (recall ou rappels). En effet, il est primordial de prédire les clients 0 qui le sont vraiment afin d'éviter des défauts de remboursement des crédits.





Parmi tous les modèles testés, la régression logistique optimisée obtient la meilleure performance à **0.7539**.

Le RandomForest obtient une performance honorable à **0.7449**.



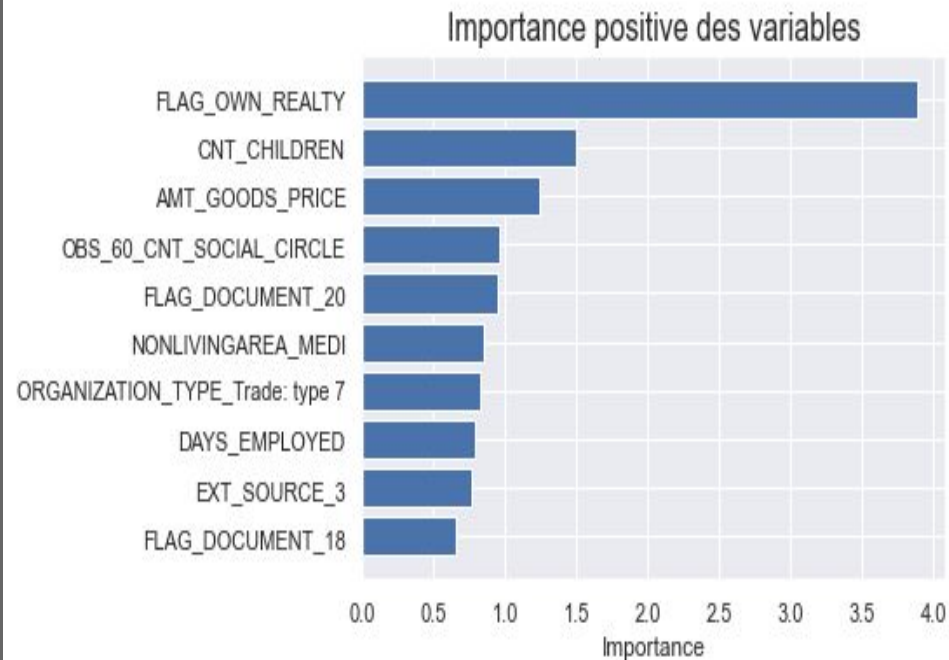
De plus, les métriques telles que la **F-mesure** et **F50-mesure** en accordant un poids de 50 pour le rappel nous indiquent que **la régression logistique optimisée** suivi du **RandomForest** sont aussi des modèles qui pénalise la non-détection.



4. INTERPRÉTABILITÉ DU MODÈLE

Importance positive des 10 premières variables :

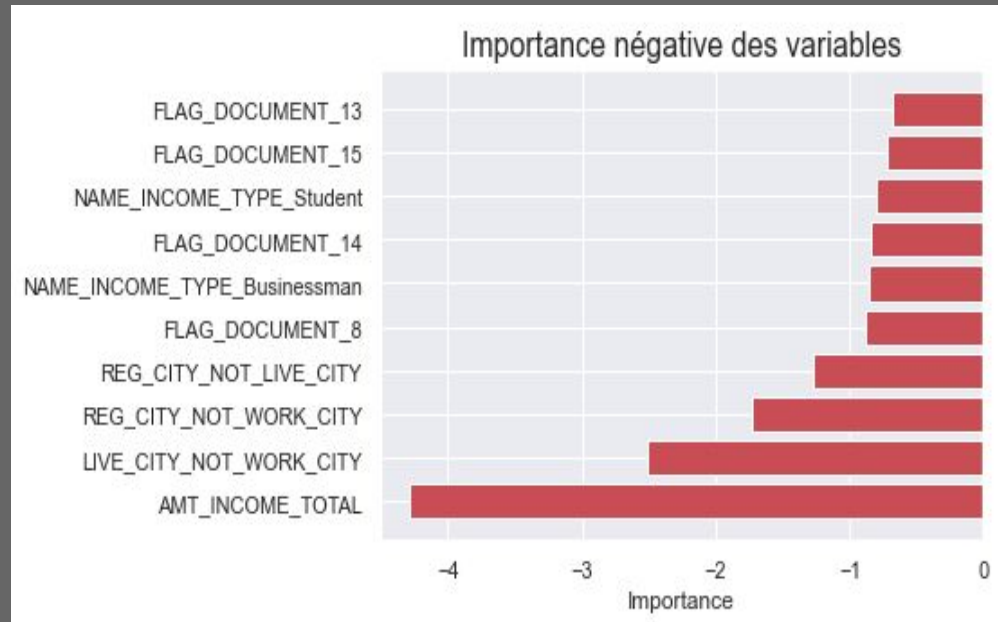
1. variable 4 (3.890693)
2. variable 5 (1.493082)
3. variable 9 (1.235686)
4. variable 79 (0.960006)
5. variable 98 (0.944256)
6. variable 75 (0.851218)
7. variable 215 (0.831719)
8. variable 12 (0.786836)
9. variable 33 (0.766221)
10. variable 96 (0.657117)



Le modèle se basant sur la régression logistique apprend sur des données telles que **FLAG_OWN_REALTY, CNT_CHILDREN, OBS_60_CNT_SOCIAL_CIRCLE, AMT_GOODS_PRICE**. En effet, ces variables vont le plus affectés le score 1 ce qui a un sens car avec plus d'enfants, un client a plus de dépenses annexes et peut être susceptible d'avoir des problème de remboursement comme dans le cas où il possède déjà un bien immobilier.

Importance négatives des 10 premières variables :

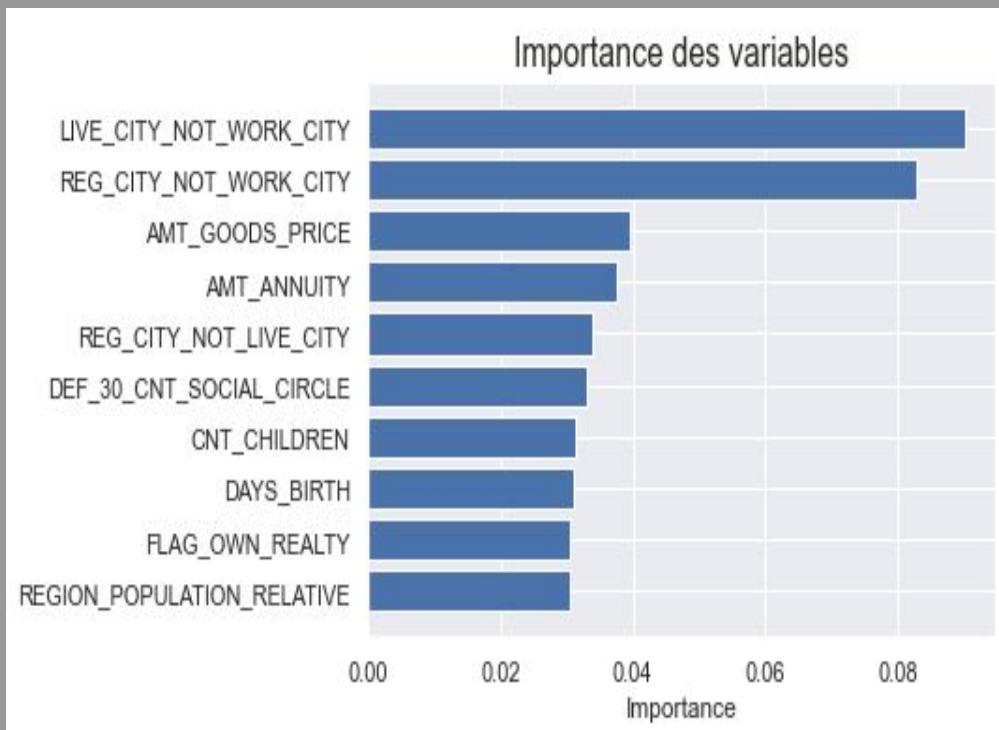
1. variable 91 (-0.681415)
2. variable 93 (-0.714800)
3. variable 120 (-0.801561)
4. variable 92 (-0.838058)
5. variable 115 (-0.855213)
6. variable 88 (-0.879302)
7. variable 28 (-1.271890)
8. variable 29 (-1.730124)
9. variable 30 (-2.509919)
10. variable 6 (-4.295019)



Des données telles **AMT_INCOME_TOTAL**, **LIVE_CITY_NOT_WORK_CITY**, **REG_CITY_NOT_WORK_CITY** vont au contraire améliorer le score qui va tendre vers 0. C'est un résultat qui fait sens car avec plus de revenus, le client pourra plus facilement rembourser son prêt. De même, si le client ne vit pas sur son lieu de travail qui est souvent localisé près des grandes agglomérations, le coût de la vie est plus faible hors des grandes villes et le client pourra rembourser plus facilement.

5. CONCLUSION

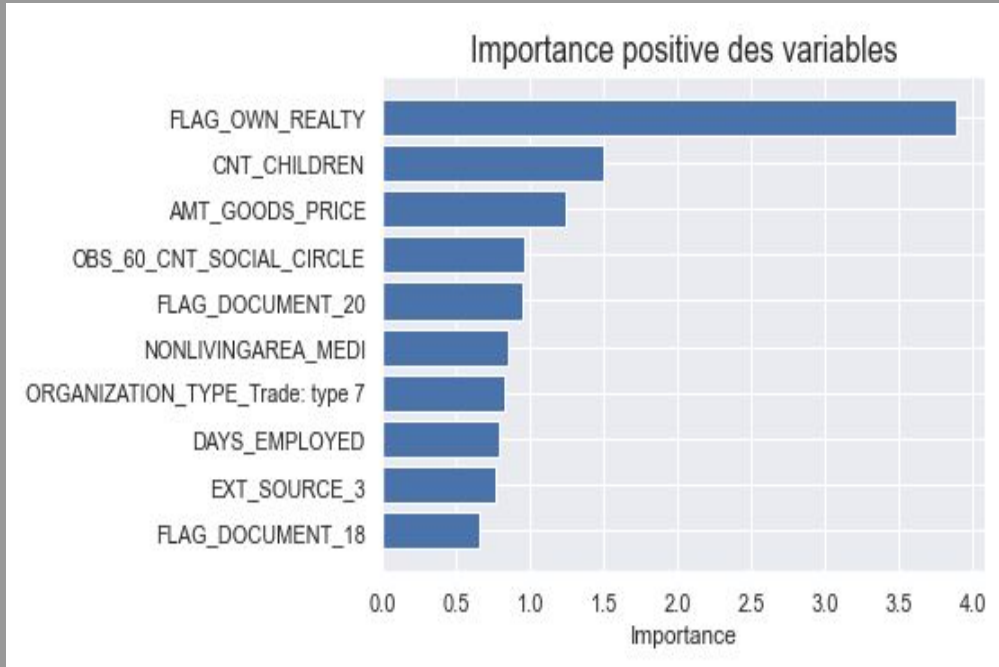




```
RFC
best params: {'criterion': 'gini',
best scores: -0.6103228646284493
Accuracy: 68.4284%
F-mesure           : 0.6797
F50-mesure         : 0.6832
Aire sous la courbe ROC: 0.7449
```

Les chargés de relation client possèdent un modèle de scoring des clients fiables à **74.49%** grâce à l'algorithme **RandomForest**. La société de prêt possède donc un outil leur permettant de minimiser les pertes financières pour les impayés (relance aux clients, frais d'avocats, frais administratifs).

De plus, la mise en évidence de l'importance des variables du modèle permet d'expliquer aux clients les raisons d'un refus de prêt.



Les chargés de relation client possèdent aussi un autre modèle de scoring des clients fiables à **75.39%** grâce à l'algorithme **de régression logistique**.

De plus, la mise en évidence de l'importance des variables du modèle permet d'augmenter la performance du modèle en fournissant au jeu d'entraînement plus de données de grandes importances.

