OPENCLASSROOMS

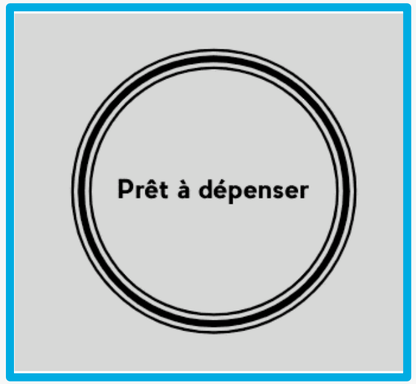
Kevin

Parcours Data Scientist

Problématique

L’entreprise

- Crédits à la consommation



Scoring

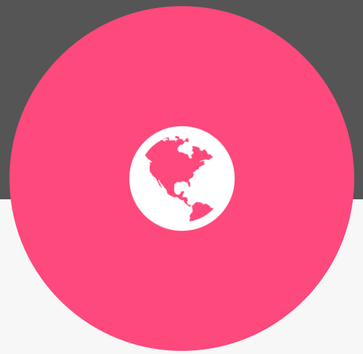
- Classification d’individus
- Algorithme de classification
- Données variées

Transparence

- Dashboard
- Utilisable par les chargés de relation client
- Satisfaction clientèle.

Jeu de données

Modélisation

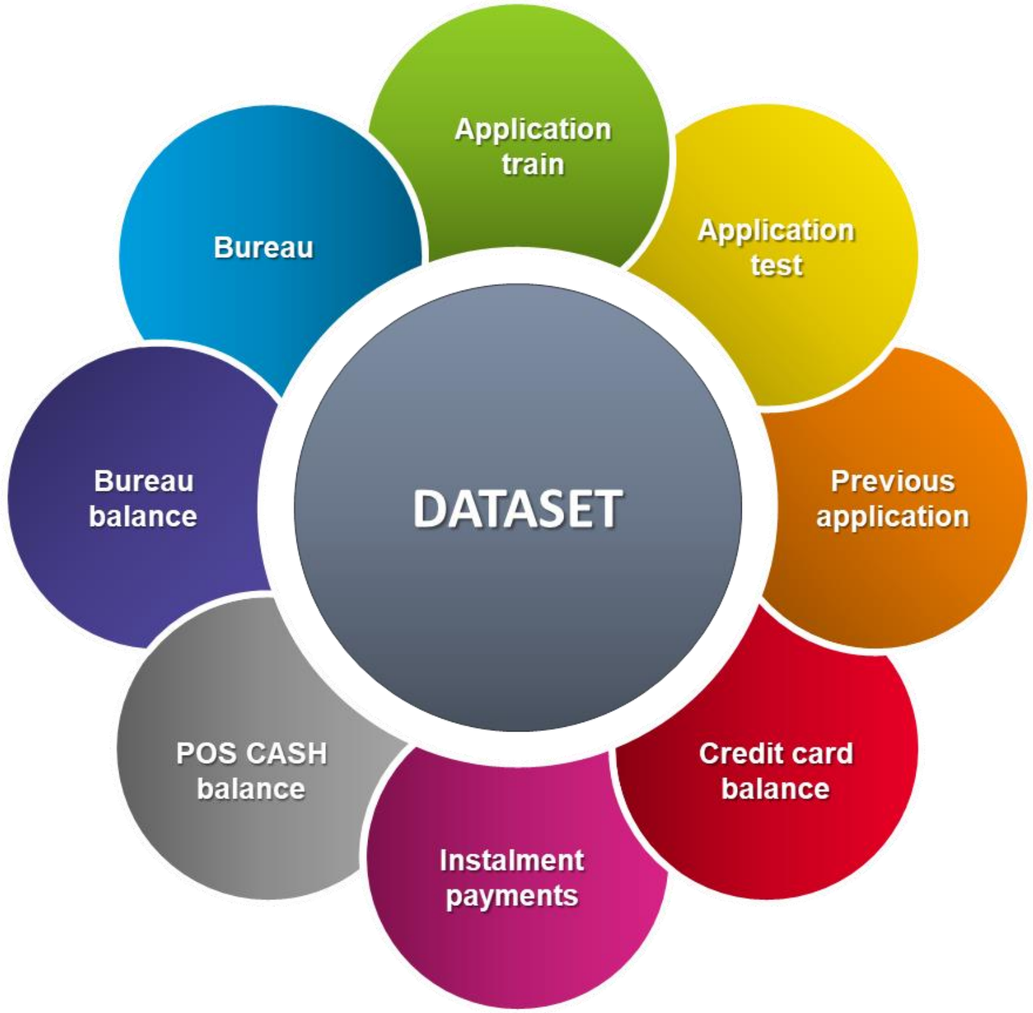


Dashboard

1

Jeu de données

Data


Application train
Application test
Bureau
Previous application
Bureau balance
DATASET
POS CASH balance
Credit card balance
Instalment payments

Fichier Data

Contenu

7 fichiers

Données

Formes très variables

Fichiers principaux

360,000 lignes

120+ variables

Fichiers historiques

1,5m à 6m de lignes

15 variables environ

Données manquantes

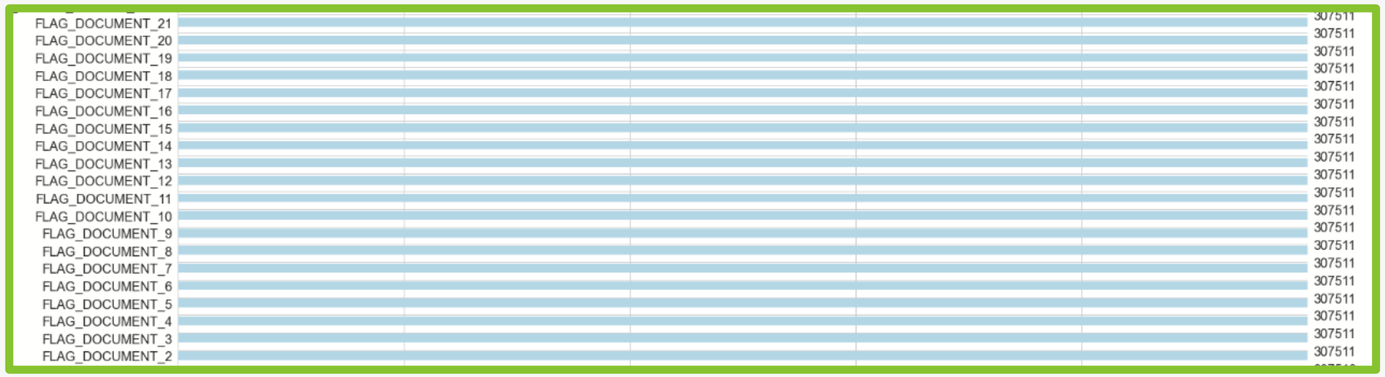
Très variables

Préparation de la data

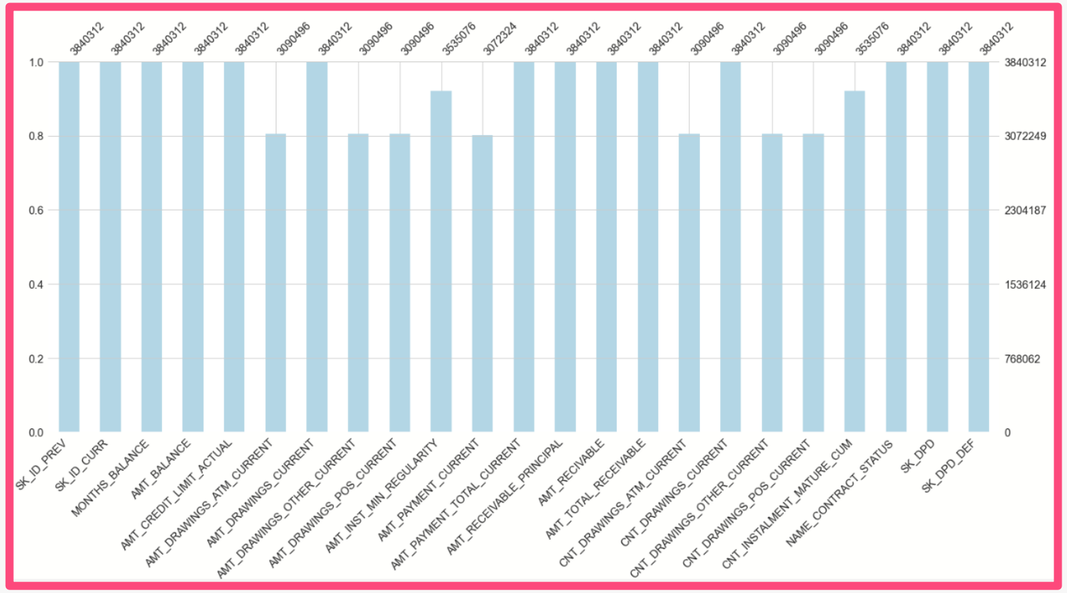
KERNEL

Données manquantes

Application train



Situation des prêts



Fichier Data

Contenu

7 fichiers

Données

Formes très variables

Fichiers principaux

360,000 lignes

120+ variables

Fichiers historiques

1,5m à 6m de lignes

15 variables environ

Données manquantes

Très variables

Préparation de la data

KERNEL

Process de la data

6 steps

Nettoyage

```
df['DAYS_EMPLOYED'].max()
✓ 0.4s
365243
```

Variables categoriques

Valeurs binaires

Valeurs avec plus de valeurs

Feature engineering

Nouvelles variables

Principalement des proportions

Aggrégations

Clients repertories plusieurs fois

Min / Max / Moyenne / Cumul

Fusion

Entrainement : 307 507 lignes - 797 variables

Test : 48 744 lignes - 797 variables

Model

LightGBM

AUC : 0,79

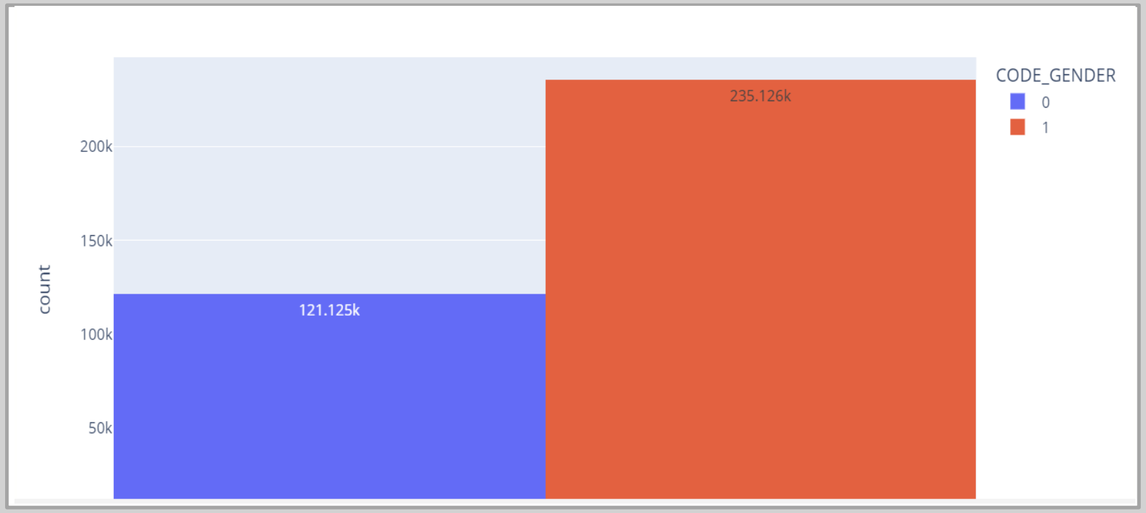
Remboursement

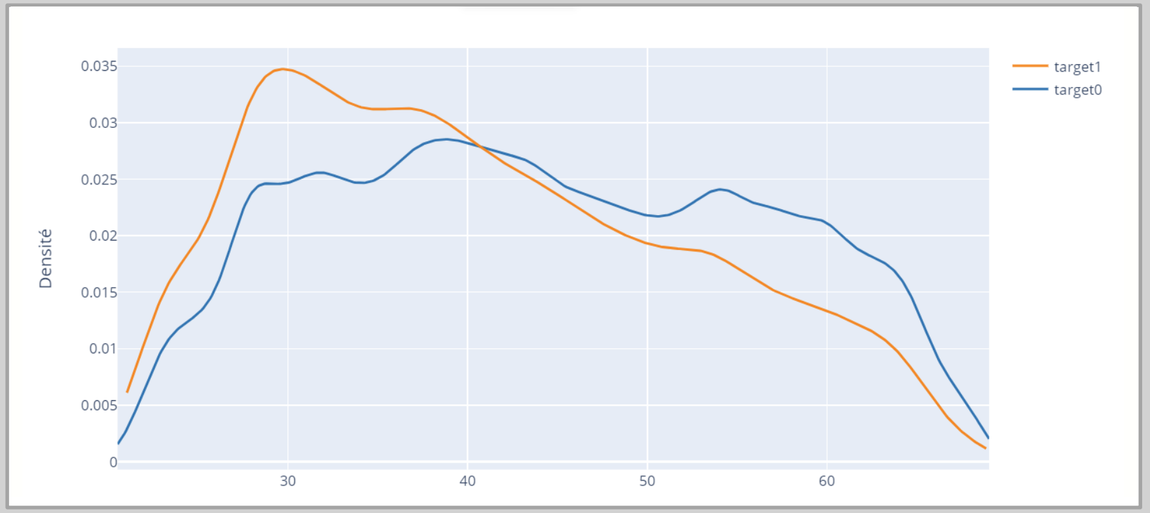
Non-remboursement

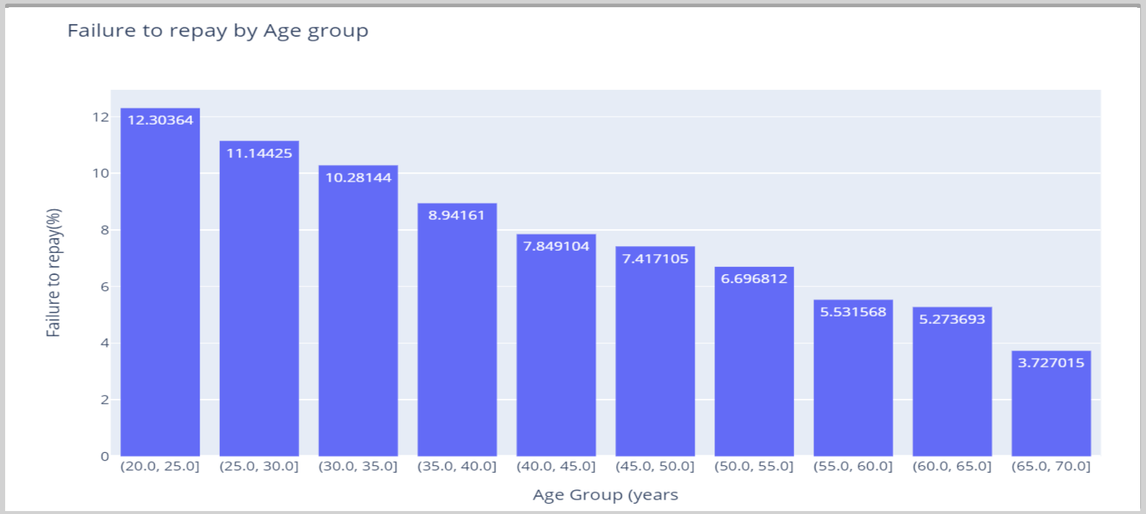
AUC

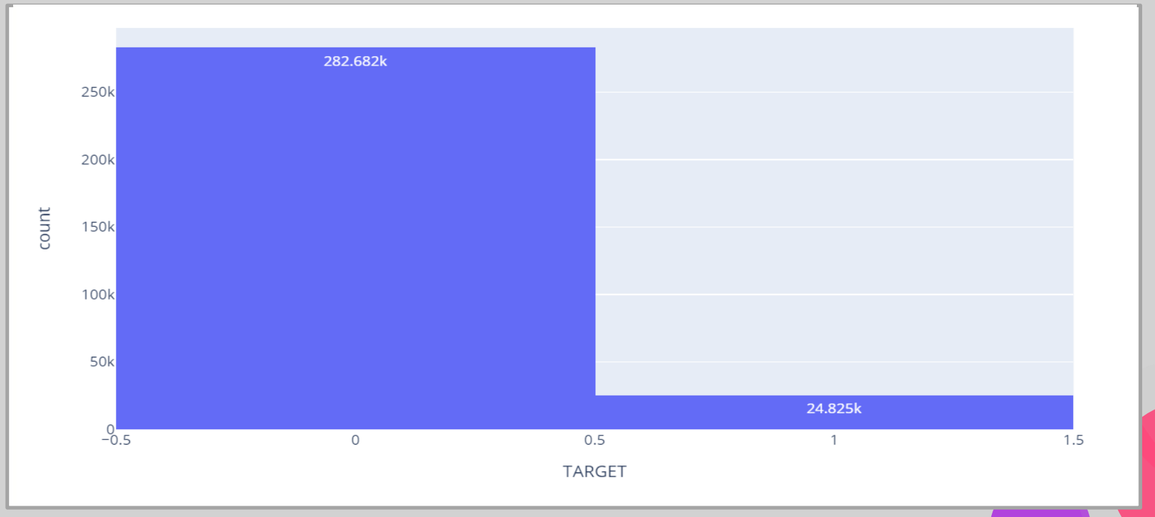
- Efficacité à distinguer des classes
- Approcher la Valeur 1
- Métrique pertinente

Analyse exploratoire









2

Modélisation

Process

Selection des features

01

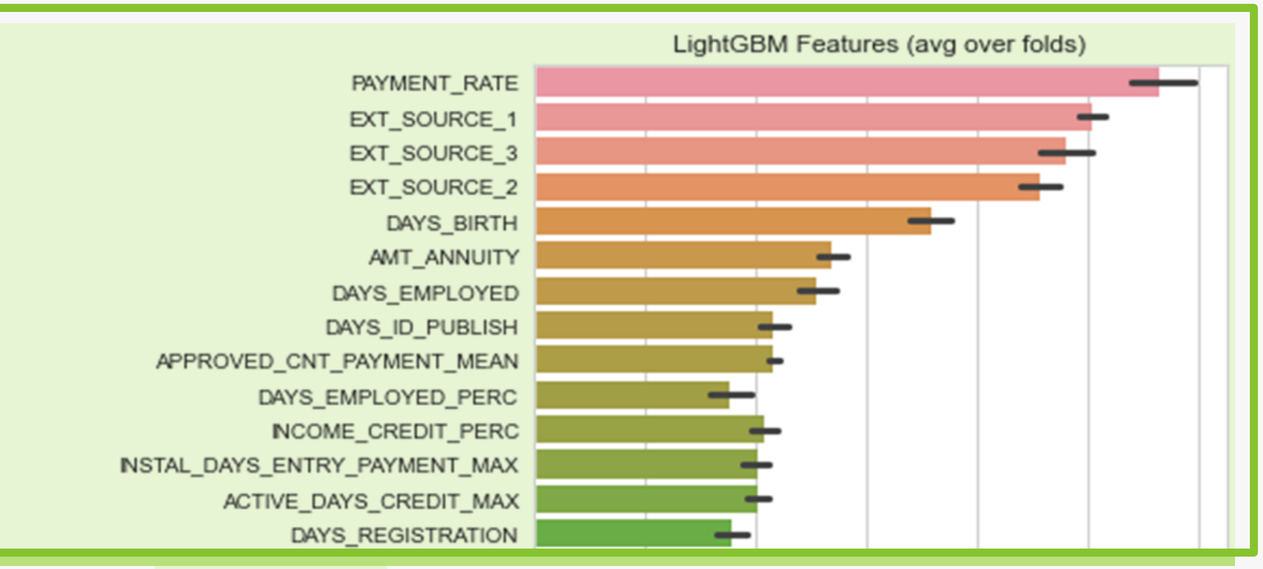
Choix de l'algorithme

02

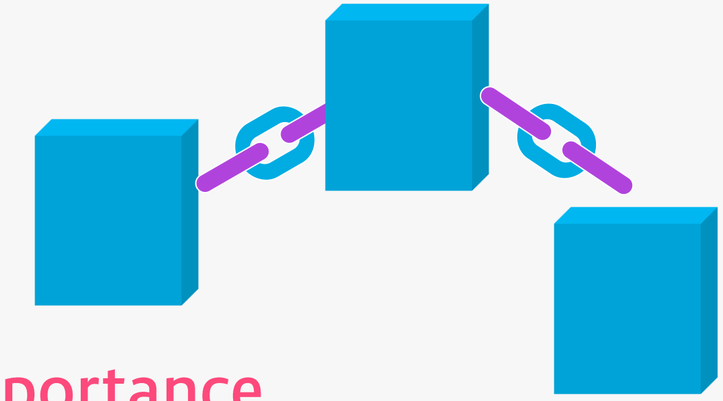
Optimisation

03

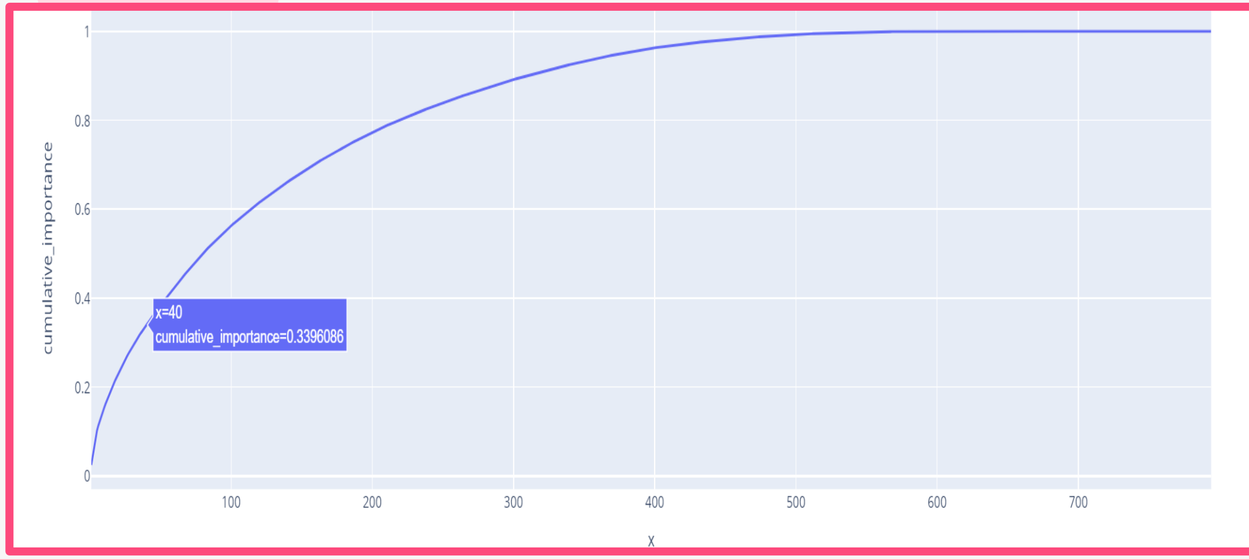
Selection des **features**



Features les plus importantes



Cumul importance



40 premières features = 1/3 de la feature importance

Selection des features

Données

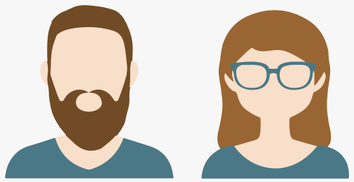
Bon équilibre
données / importance

Utilisation

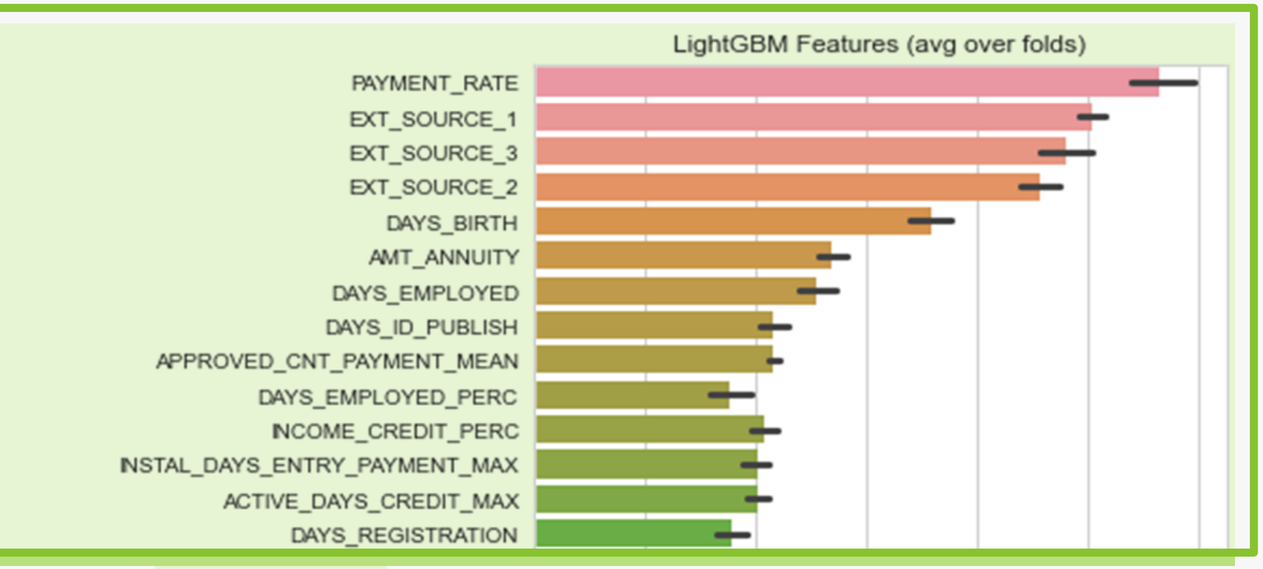
Simplicité et rapidité

Transparence

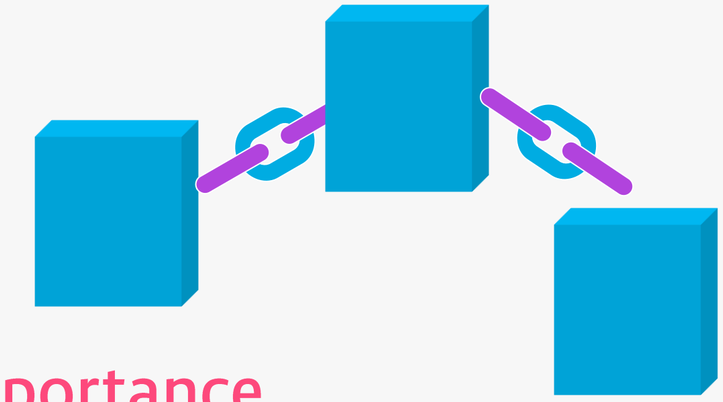
Discriminations



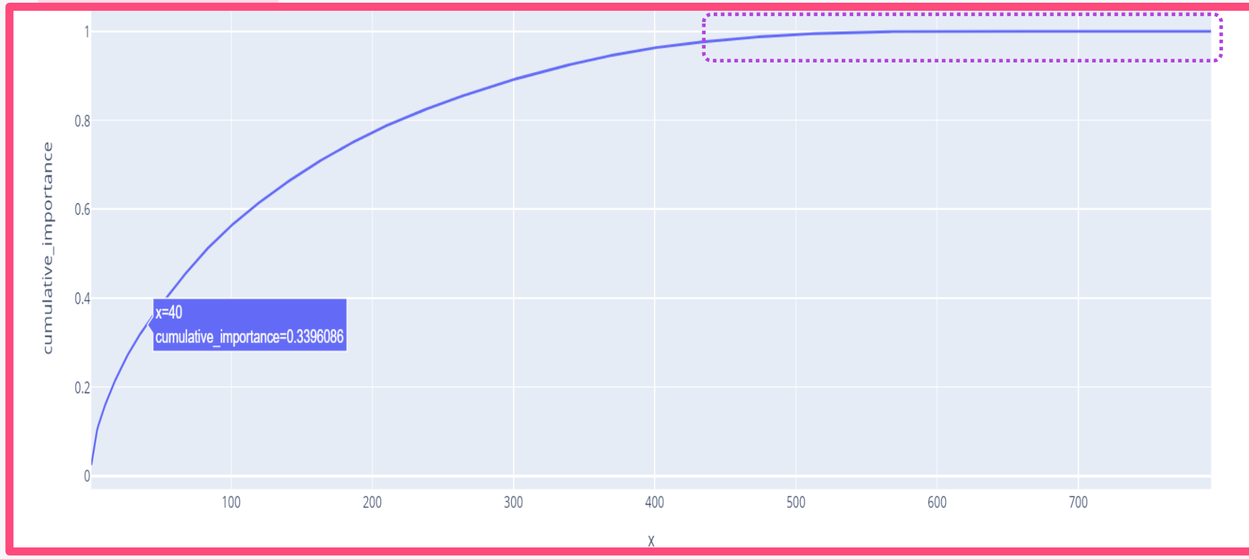
Selection des **features**



Features les plus importantes



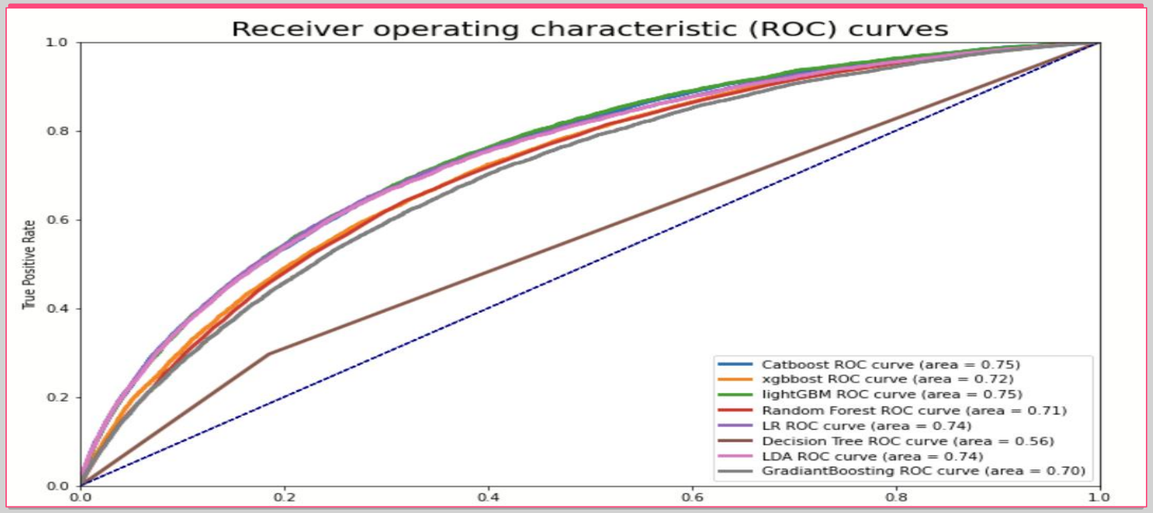
Cumul importance

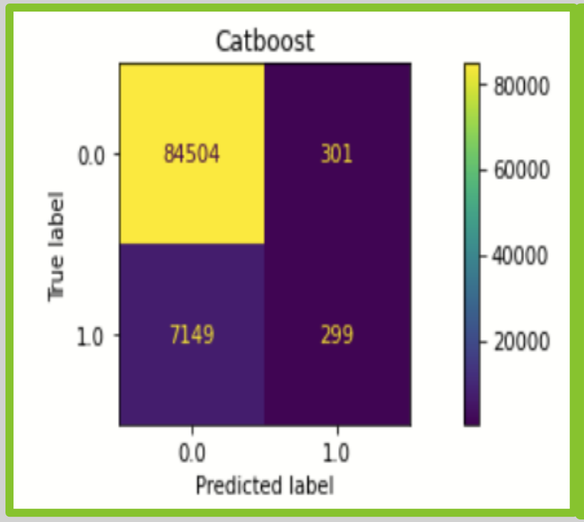


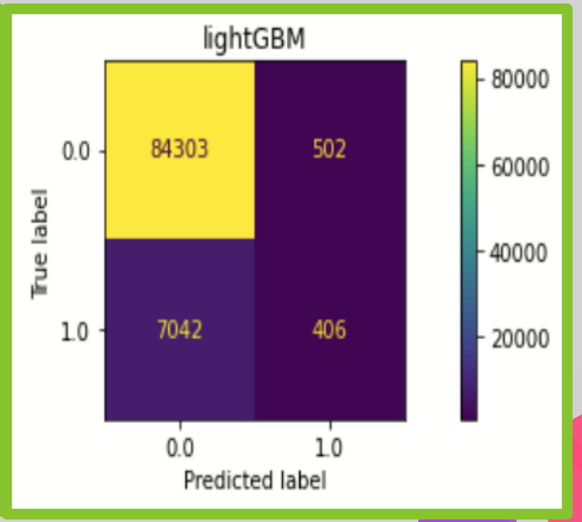
40 premières features = 1/3 de la feature importance

Tests d’algorithmes

	Classifier	Accuracy	ROC_AUC Test	ROC_AUC Train	Recall	Precision	F1
2	lightGBM	91.533067	0.749213	0.980652	0.075054	0.377448	0.125210
0	Catboost	91.646884	0.746679	0.841128	0.058405	0.385638	0.101446
4	LR	76.472310	0.744128	0.745463	0.567132	0.186038	0.280171
6	LDA	76.317301	0.741761	0.743622	0.562701	0.183961	0.277274
1	xgbbost	90.716833	0.719652	0.801418	0.090494	0.273539	0.135997
3	Random Forest	88.507691	0.714885	1.000000	0.176155	0.227068	0.198397
7	GradiantBoosting	82.620619	0.702585	0.709653	0.338883	0.185139	0.239457
5	Decision Tree	77.292879	0.555738	1.000000	0.296724	0.123326	0.174235







Process

4 steps

Pipeline Une ou plusieurs étapes

Découpage du dataset 70% entrainement – 30% test

Metrique d'évaluation (Train/test) Accuracy – F1 Score – F1 Beta score - AUC

Métrique d'évaluation (Test)

Matrice de confusion

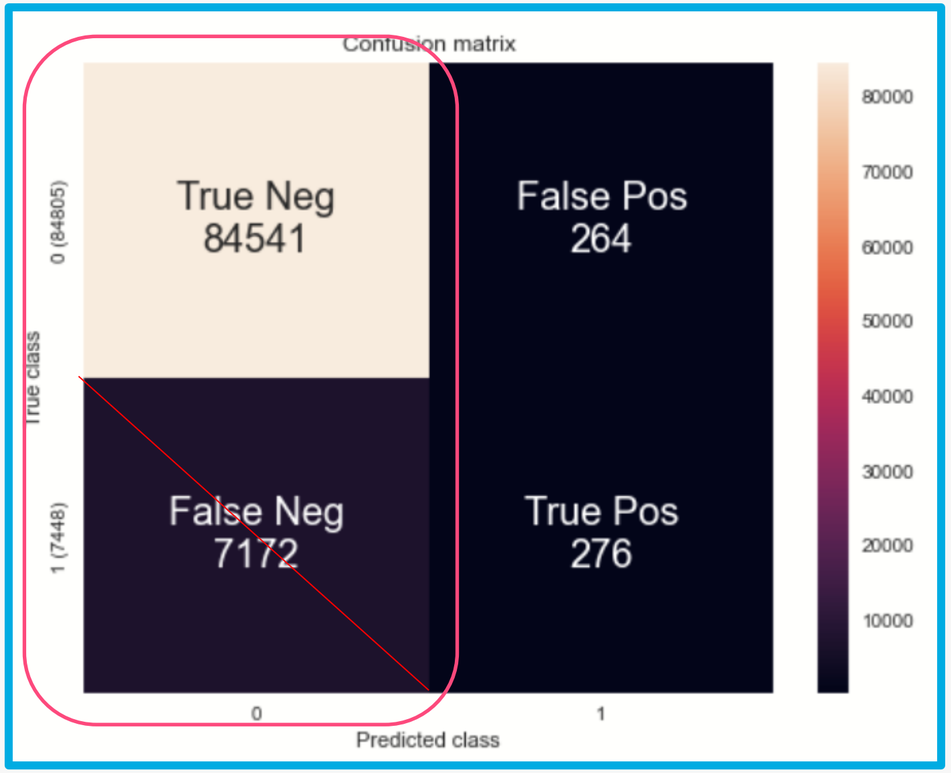
Classification report

Recall plot

ROC – AUC Curve

Objectif

Tout le monde rembourse



Objectifs :

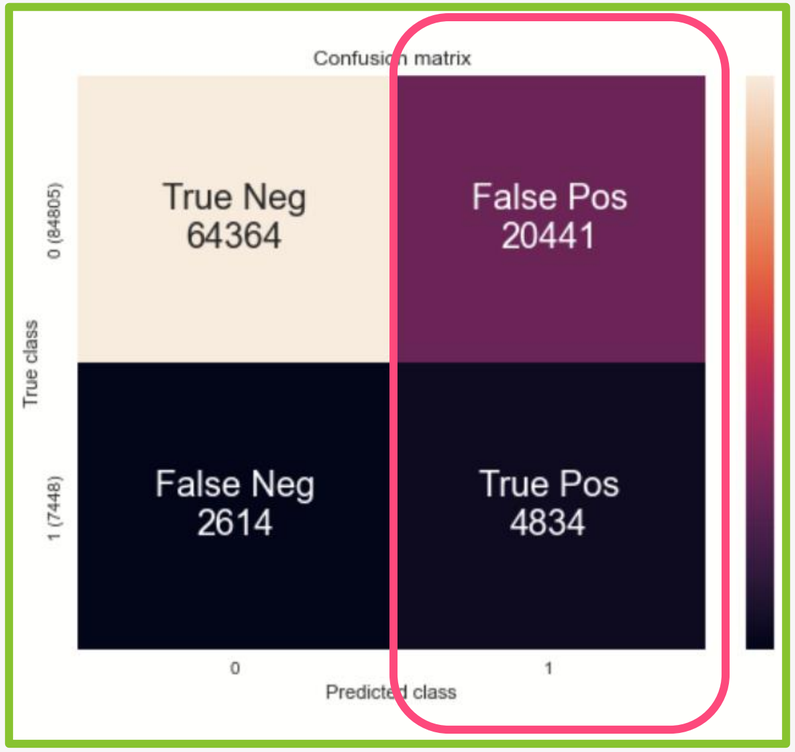
- Eviter les faux negatifs en priorité
- Maximiser les True negatifs

Application du process

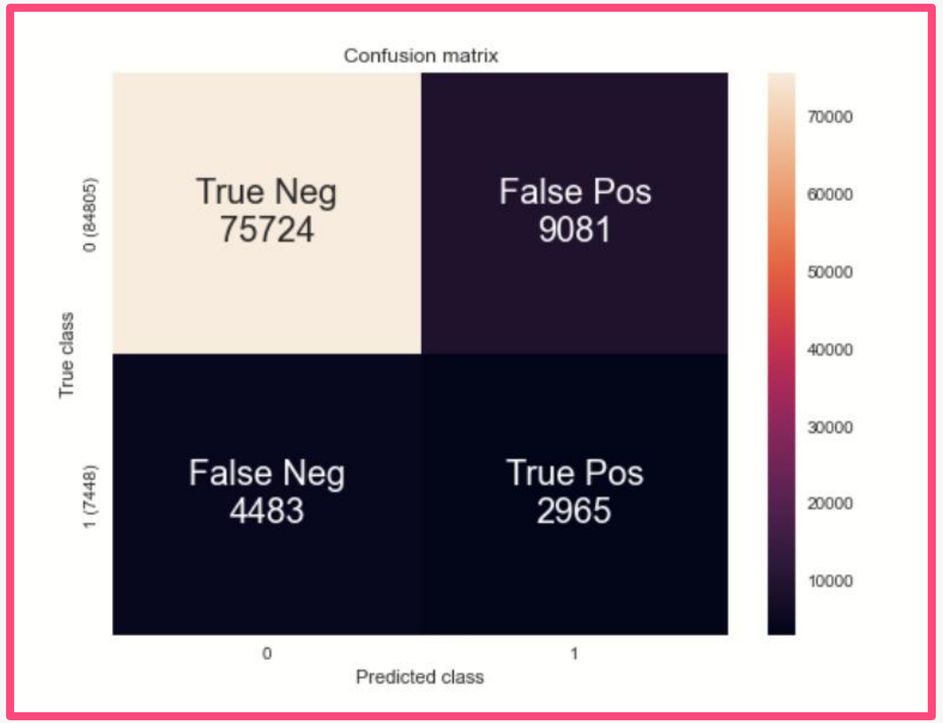
Entrainement : 307 507 lignes / Test : 48744 lignes

Sans modification	Uniquement LightGBM	0,77
Balanced	Uniquement LightGBM avec équilibrage du poids des classes	0,77
Imputation, Standardisation et Balanced	SimpleImputer / IterativeImputer + StandardScaler + Equilibrage du poids des classes	0,76

Balanced



Pipeline



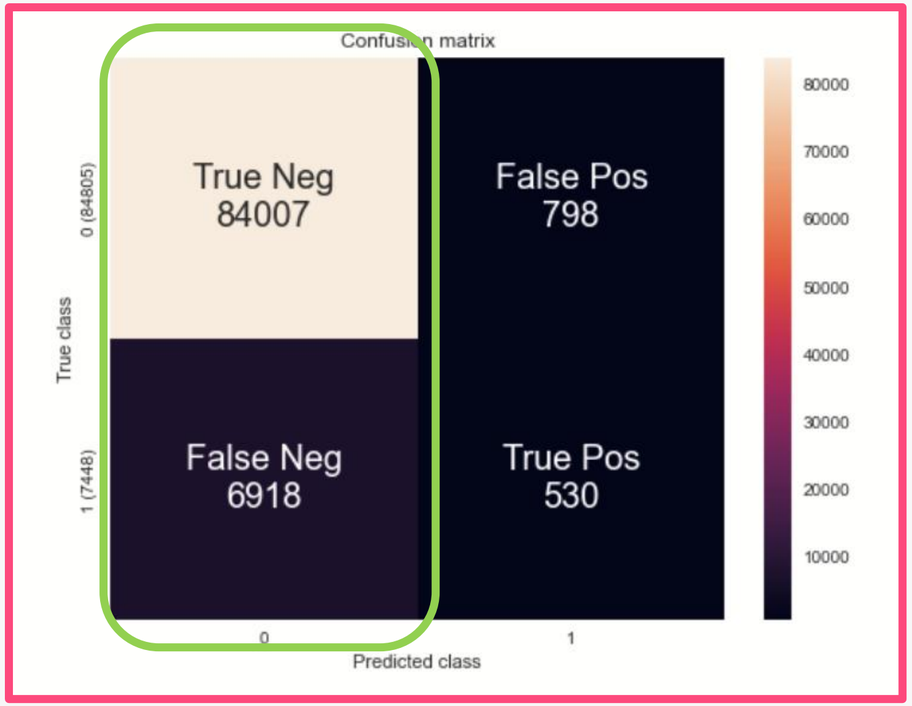
Application du process

Entrainement : 307 507 lignes / Test : 48744 lignes

Sans modification	Uniquement LightGBM	0,77
Balanced	Uniquement LightGBM avec équilibrage du poids des classes	0,77
Imputation, Standardisation et Balanced	SimpleImputer / IterativeImputer + StandardScaler + Equilibrage du poids des classes	0,76
Imputation et SMOTE	SimpleImputer / IterativeImputer + Création d'agents fictifs (SMOTE)	0,76

Imputation et SMOTE.

Peu efficace…

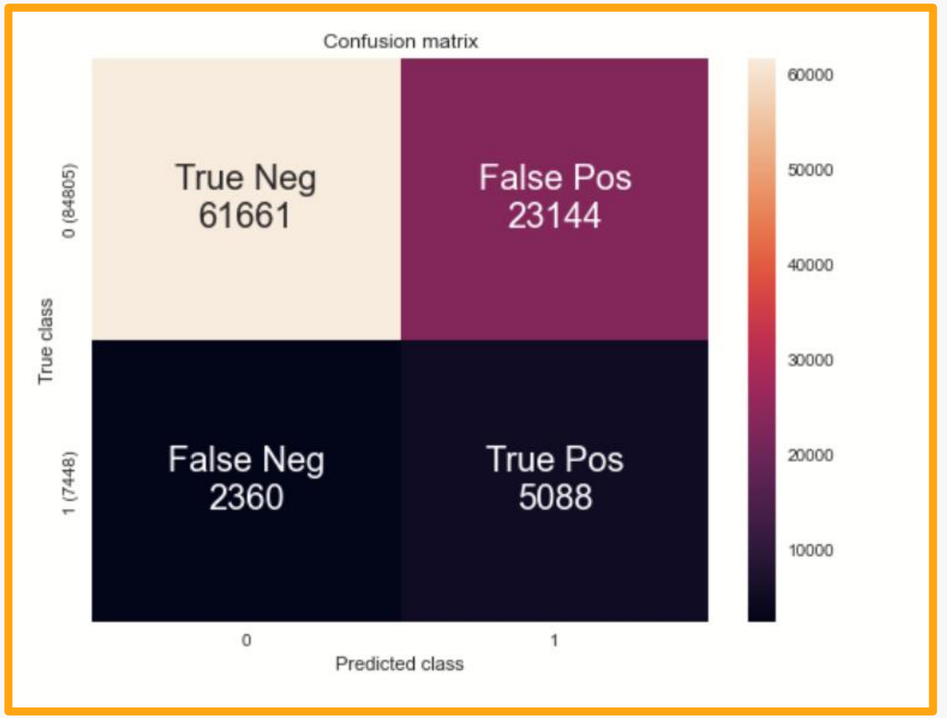


Application du process

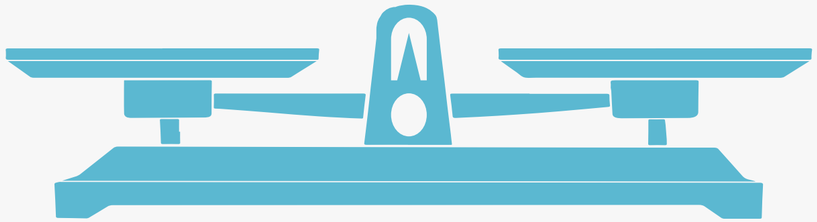
Entrainement : 307 507 lignes / Test : 48744 lignes

Sans modification	Uniquement LightGBM	0,77
Balanced	Uniquement LightGBM avec équilibrage du poids des classes	0,77
Imputation, Standardisation et Balanced	SimpleImputer / IterativeImputer + StandardScaler + Equilibrage du poids des classes	0,76
Imputation et SMOTE	SimpleImputer / IterativeImputer + Création d'agents fictifs (SMOTE)	0,76
Optimisation	GridSearchCV / RandomizedSearchCV	

Optimisation



Metrique personnalisée



True négatif

Personne qui rembourse et bien identifié

Faux positif

Personne qui rembourse et mal identifié

+1	0
-10	0

Faux négatif

Personne qui ne rembourse pas et mal identifié

True positif

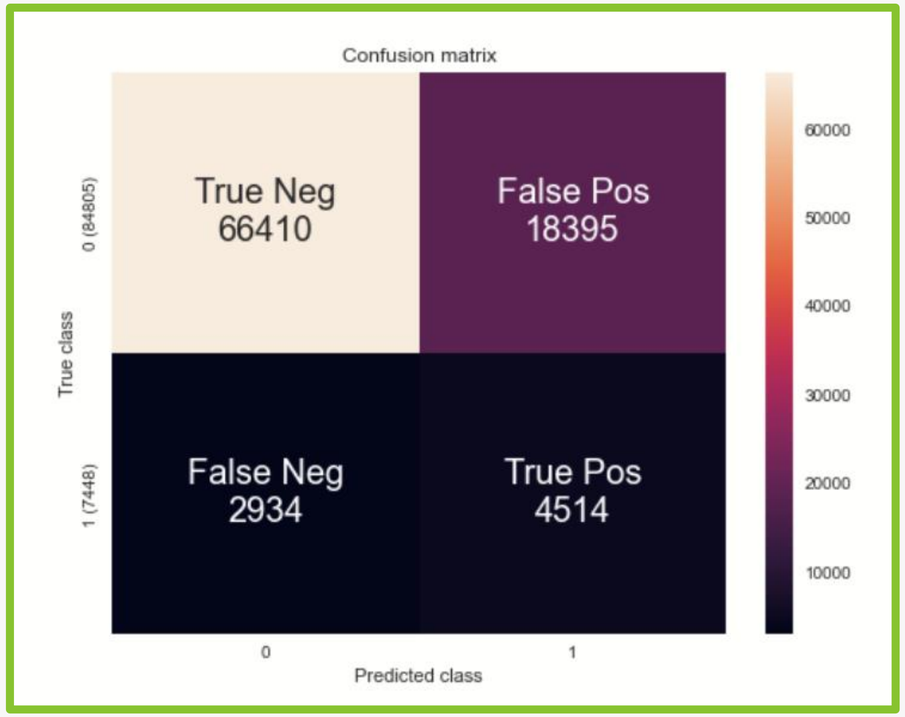
Personne qui ne rembourse pas, et bien identifié

	Classifier	Accuracy	ROC_AUC Test	ROC_AUC Train	Recall	Precision	F1	Scoring
0	Catboost(autoscale)	76.879885	0.765014	0.905867	0.606069	0.197040	0.297394	0.401830
1	Catboost(scale_weight)	76.638158	0.764079	0.905538	0.602175	0.194375	0.293886	0.396583
2	xgbbost	75.187799	0.755897	0.884264	0.609694	0.185166	0.284061	0.387543
5	LR	68.722968	0.744580	0.746709	0.674409	0.159704	0.258252	0.369918
3	lightGBM	85.296955	0.757173	0.995310	0.398093	0.246140	0.304196	0.334883
6	Decision Tree	86.157632	0.540146	1.000000	0.156821	0.152520	0.154641	0.168179
7	LDA	91.905954	0.740854	0.743109	0.018931	0.468439	0.036392	0.125470
8	GradiantBoosting	91.953649	0.762051	0.773726	0.017320	0.553648	0.033589	0.124776
4	Random Forest	91.911374	0.734059	1.000000	0.005371	0.425532	0.010607	0.115671

Choix de l’algorithme définitif

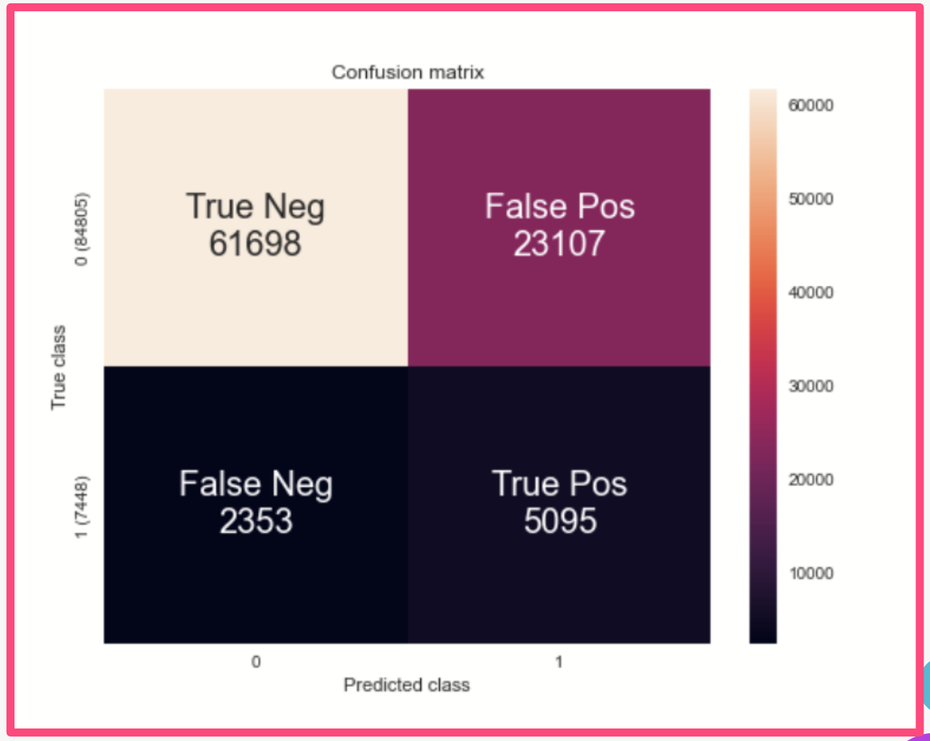
Choix de l'algorithme

Catboost

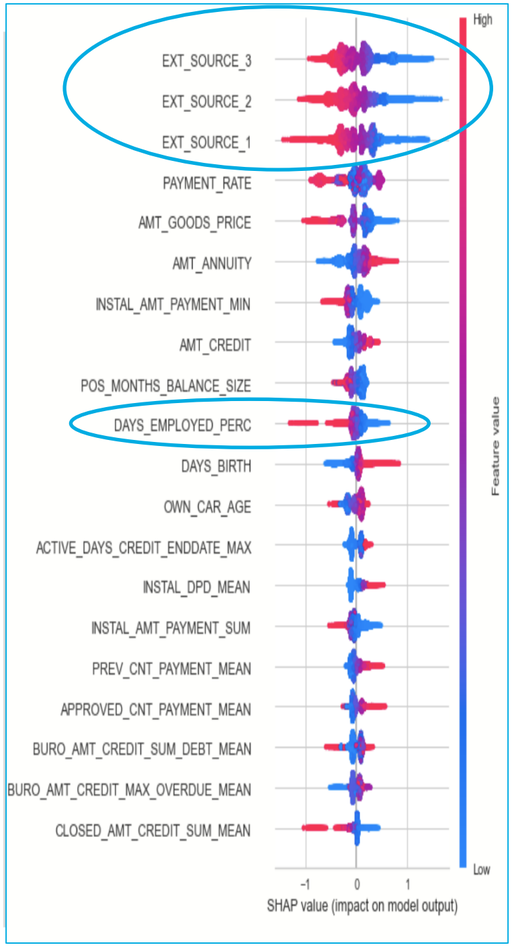


LightGBM

Private Score	Public Score
0.76422	0.76218



Interprétabilité globale



3

Dashboard

Process

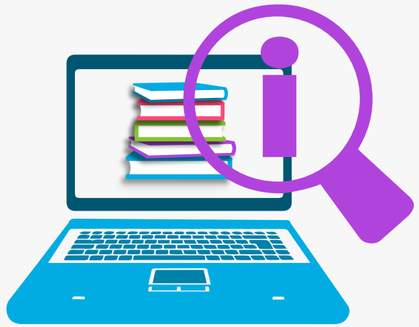
Backend

- FastApi


```
{
  "PAYMENT_RATE": 0.0607492667810303,
  "EXT_SOURCE_1": 0.0830369673913225,
  "EXT_SOURCE_3": 0.1393757800997895,
  "EXT_SOURCE_2": 0.2629485927471776,
  "DAYS_BIRTH": 9461,
  "AMT_ANNUITY": 24700.5,
  "DAYS_EMPLOYED": -637,
  "APPROVED_CNT_PAYMENT_MEAN": 24,
  "DAYS_ID_PUBLISH": -2120,
  "INCOME_CREDIT_PERC": 0.1219777777777777,
  "ACTIVE_DAYS_CREDIT_MAX": -103,
  "INSTAL_DAYS_ENTRY_PAYMENT_MAX": -49,
  "INSTAL_DPD_MEAN": 0,
  "DAYS_REGISTRATION": -3648,
  "DAYS_EMPLOYED_PERC": 0.0673290349857309,
  "ACTIVE_DAYS_CREDIT_ENDDATE_MIN": 780,
  "AMT_CREDIT": 406597.5,
  "PREV_CNT_PAYMENT_MEAN": 24,
  "AMT_GOODS_PRICE": 351000,
  "INSTAL_AMT_PAYMENT_SUM": 219625.695,
  "REGION_POPULATION_RELATIVE": 0.018801,
  "INSTAL_DBD_SUM": 388,
  "DAYS_LAST_PHONE_CHANGE": -1134,
  "BURO_AMT_CREDIT_MAX_OVERDUE_MEAN": 1681.029,
  "CLOSED_DAYS_CREDIT_MAX": -476,
  "OWN_CAR_AGE": "Indisp",
  "CLOSED_DAYS_CREDIT_ENDDATE_MAX": 85,
```

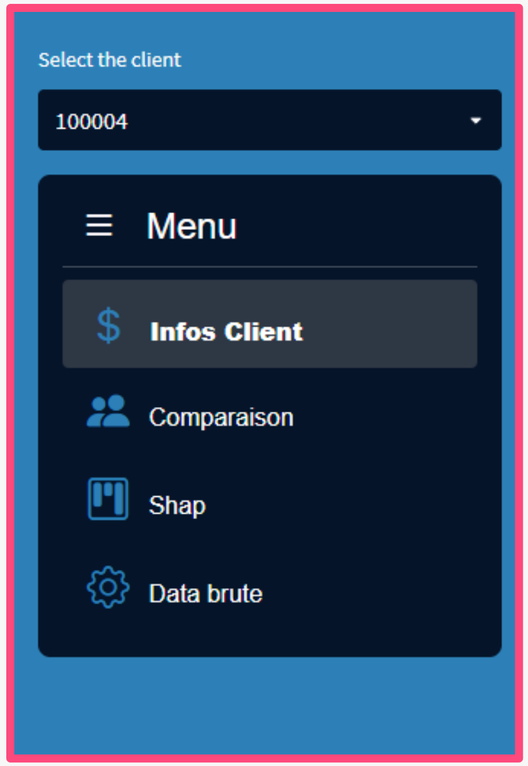


```
{
  "result": 0.84
}
```

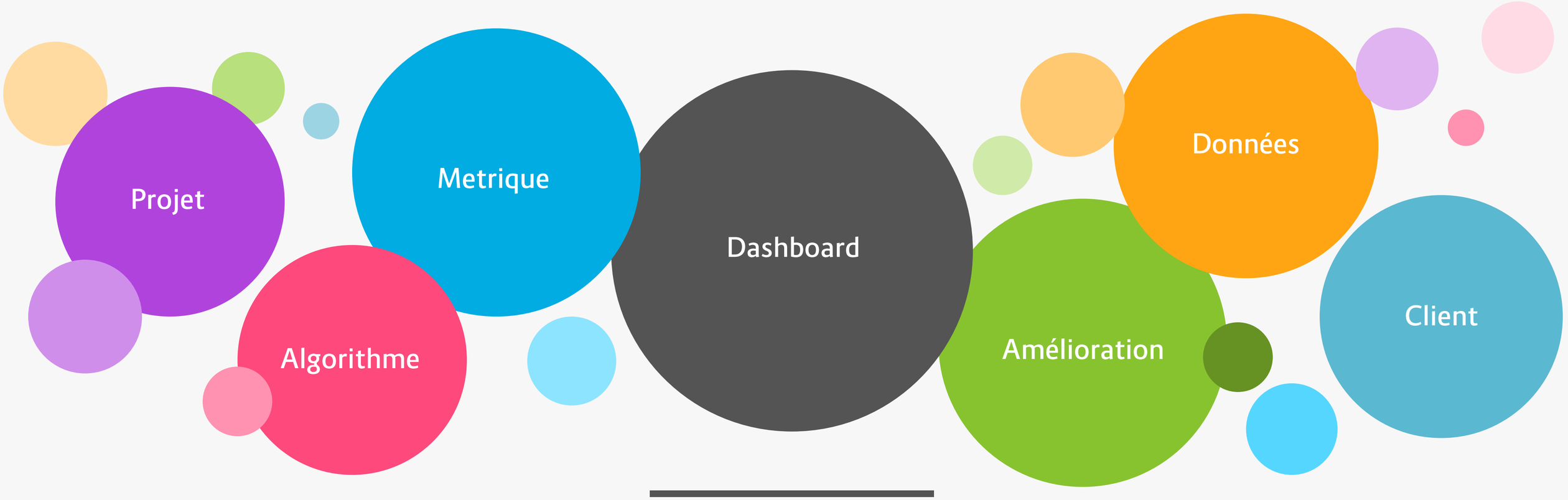



Frontend

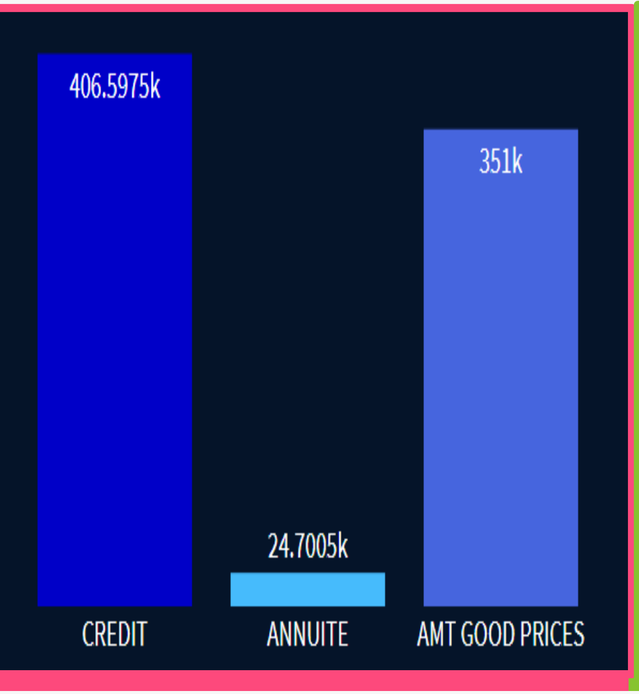
- Streamlit

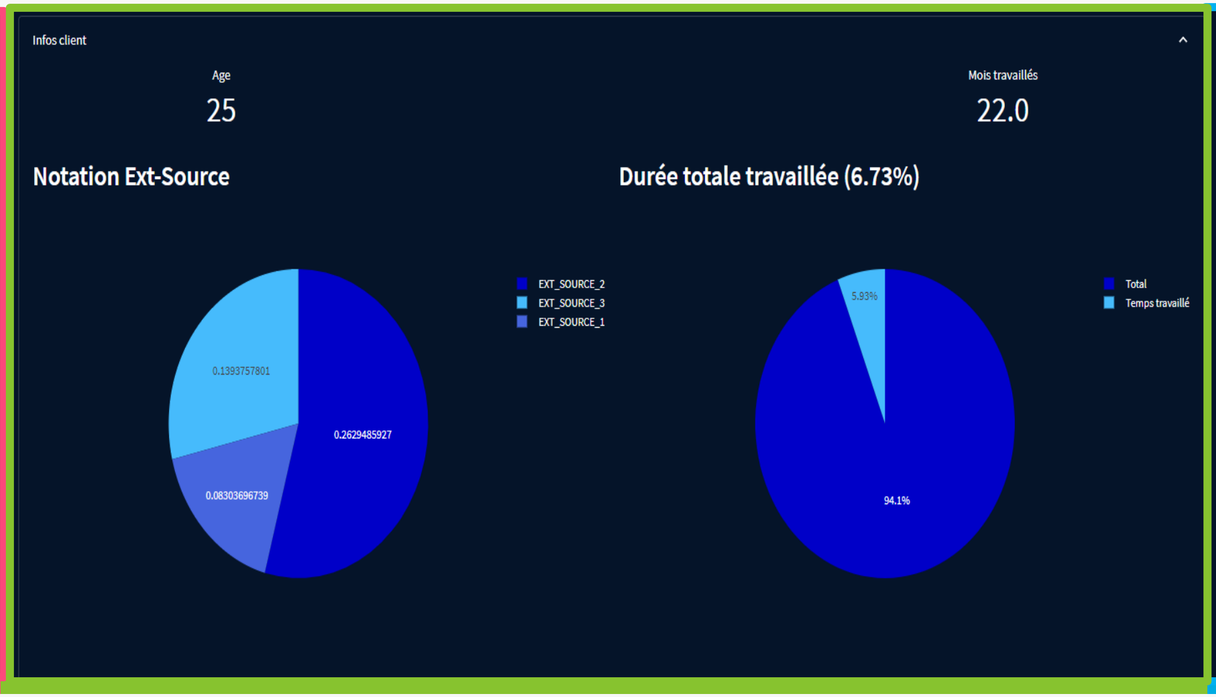


Conclusion


Projet
Metrique
Algorithme
Dashboard
Amélioration
Données
Client

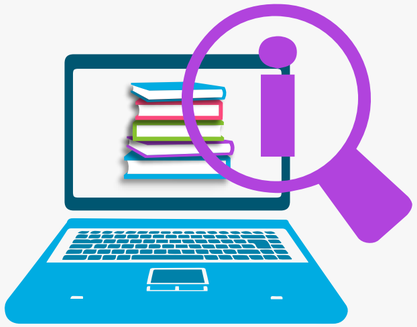
Page **Principale**





Comparaison client

Loop with icons and descriptions



EXT_SOURCE

- EXT_SOURCE_1
- EXT_SOURCE_2

CREDIT

- MONTANT CREDIT
- MONTANT ANNUITE
- GOODS PRICE

PAYMENT RATE

INFOS PERSONNELLES

- Age
- Nombre de jours travaillés

...

Shap


= −0.611
EXT_SOURCE_2 −1
EXT_SOURCE_3 +0.85
EXT_SOURCE_1 −0.36
INSTAL_AMT_PAYMENT_SUM +0.21
POS_MONTHS_BALANCE_SIZE +0.14
ACTIVE_DAYS_CREDIT_UPDATE_MEAN −0.13
OWN_CAR_AGE −0.11
AMT_CREDIT +0.11
PREV_CNT_PAYMENT_MEAN −0.11
INSTAL_DPD_MEAN +0.1
INSTAL_AMT_PAYMENT_MIN −0.1
ACTIVE_DAYS_CREDIT_MAX +0.09
BURO_DAYS_CREDIT_MAX +0.09
AMT_GOODS_PRICE +0.09
AMT_ANNUITY −0.08
PAYMENT_RATE −0.08
DAYS_BIRTH +0.08
APPROVED_CNT_PAYMENT_MEAN −0.08
DAYS_LAST_PHONE_CHANGE −0.08
ACTIVE_DAYS_CREDIT_ENDDATE_MAX +0.08
DAYS_ID_PUBLISH +0.07
BURO_AMT_CREDIT_MAX_OVERDUE_MEAN +0.07
INSTAL_DAYS_ENTRY_PAYMENT_MAX −0.07
INSTAL_DBD_SUM +0.06
BURO_AMT_CREDIT_SUM_DEBT_MEAN +0.06
ACTIVE_DAYS_CREDIT_MEAN +0.06
INCOME_CREDIT_PERC −0.05
APPROVED_DAYS_DECISION_MAX −0.05
REGION_POPULATION_RELATIVE +0.05
ACTIVE_AMT_CREDIT_SUM_SUM −0.04
INSTAL_DBD_MAX +0.04
CLOSED_DAYS_CREDIT_MAX −0.04
ACTIVE_DAYS_CREDIT_ENDDATE_MEAN +0.02
CLOSED_AMT_CREDIT_SUM_MEAN +0.02
DAYS_REGISTRATION +0.02
CLOSED_DAYS_CREDIT_ENDDATE_MAX −0.02
ACTIVE_DAYS_CREDIT_ENDDATE_MIN +0.01
BURO_DAYS_CREDIT_ENDDATE_MAX +0.01
DAYS_EMPLOYED −0.01
DAYS_EMPLOYED_PERC −0
−0.6 −0.4 −0.2 0.0 0.2 0.4
= −0.555

EXT_SOURCE_2 (0.263)
EXT_SOURCE_3 (0.139)
EXT_SOURCE_1 (0.083)
INSTAL_AMT_PAYMENT_SUM (219,625.695)
POS_MONTHS_BALANCE_SIZE (19)
ACTIVE_DAYS_CREDIT_UPDATE_MEAN (-15.5)
OWN_CAR_AGE (Indisp)
AMT_CREDIT (406,597.5)
PREV_CNT_PAYMENT_MEAN (24)
INSTAL_DPD_MEAN (0)
INSTAL_AMT_PAYMENT_MIN (9,251.775)
ACTIVE_DAYS_CREDIT_MAX (-103)
BURO_DAYS_CREDIT_MAX (-103)
AMT_GOODS_PRICE (351,000)
AMT_ANNUITY (24,700.5)
PAYMENT_RATE (0.061)
DAYS_BIRTH (9,461)
APPROVED_CNT_PAYMENT_MEAN (24)
DAYS_LAST_PHONE_CHANGE (-1,134)
ACTIVE_DAYS_CREDIT_ENDDATE_MAX (780)
Model output value