

Sommaire



Contexte



Présentation du jeu de données



Traitement du jeu de données



Modélisation



Explication du modèle



Conclusion



Contexte



Contexte



CRÉDIT À LA CONSOMMATION



FACILEMENT INTERPRÉTABLE



IMPORTANCE DES VARIABLES



Présentation du jeu de données



Analyse de la forme

307.511 Lignes et 122 colonnes

Répartition « Target » désequilibrée

	Nombre de valeur	Pourcentage
0	282686.0	91.93
1	24825.0	8.07



Analyse de la forme

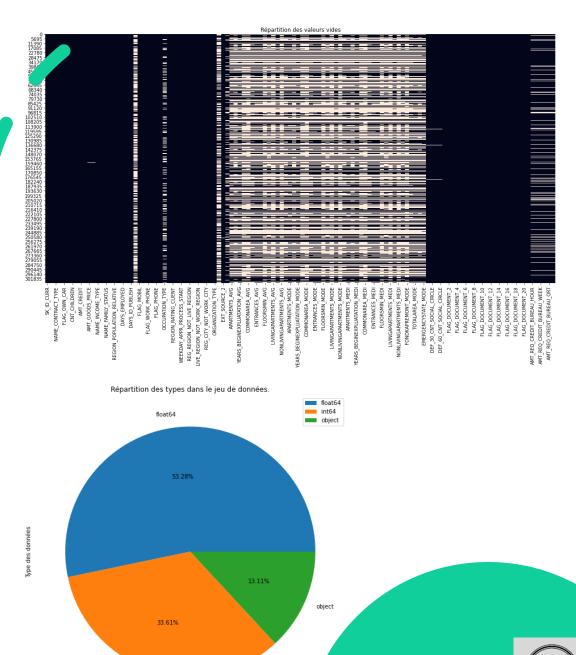
307.511 Lignes et 122 colonnes Répartition « Target » désequilibrée

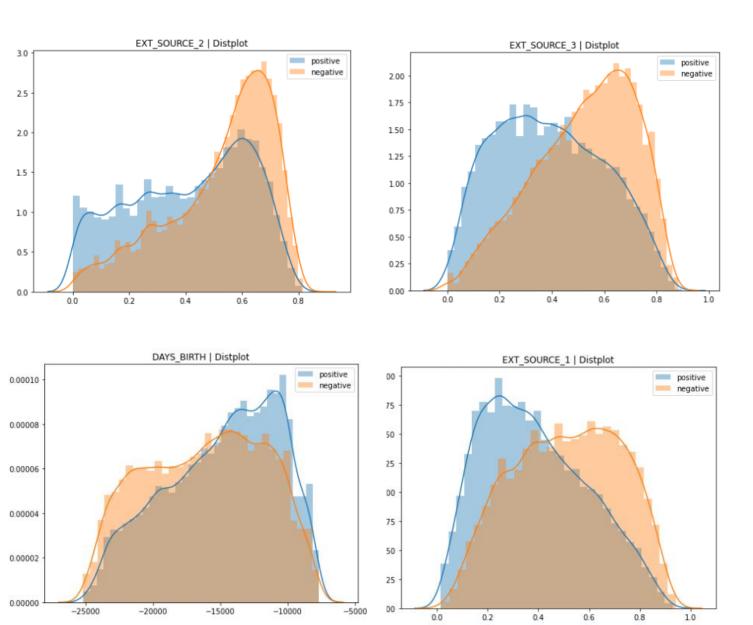
Nombre de valeur		Pourcentage
remboursé	282686.0	91.93
non-remboursé	24825.0	8.07



Analyse de la forme (suite)

Répartitions des valeurs vides Répartition des types de données



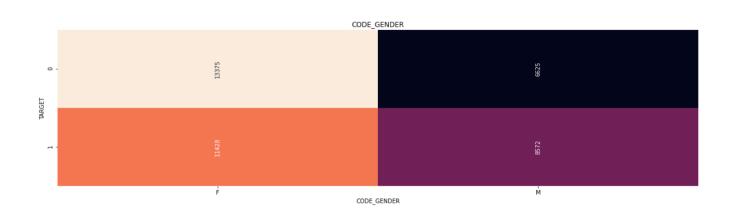


Analyse de fond



Valeurs numériques







Analyse de fond





Valeurs numériques Valeurs catégorielles



Carte de chaleur des chiffres à virgule flottantes SK_ID_CURR : CNT_CHILDREN : AMT_CREDIT : AMT GOODS PRICE DAYS BIRTH DAYS REGISTRATION OWN CAR AGE FLAG EMP PHONE FLAG_CONT_MOBILE FLAG EMAIL REGION RATING CLIENT HOUR APPR PROCESS START REG REGION NOT WORK REGION REG CITY NOT LIVE CITY LIVE_CITY_NOT_WORK_CITY -EXT_SOURCE_2 APARTMENTS_AVG YEARS_BEGINEXPLUATATION_AVG ENTRANCES AVG FLOORSMIN AVG LIVINGAPARTMENTS AVG NONLIVINGAPARTMENTS AVG APARTMENTS MODE YEARS BEGINEXPLUATATION MODE COMMONAREA MODE ENTRANCES_MODE FLOORSMIN MODE LIVINGAPARTMENTS MODE NONLIVINGAPARTMENTS MODE APARTMENTS MEDI YEARS BEGINEXPLUATATION MEDI COMMONAREA MEDI ENTRANCES MEDI FLOORSMIN MEDI LIVINGAPARTMENTS MEDI NONLIVINGAPARTMENTS_MEDI TOTALAREA MODE DEF_30_CNT_SOCIAL_CIRCLE DEF 60 CNT SOCIAL CIRCLE FLAG DOCUMENT 2 FLAG DOCUMENT 4 FLAG DOCUMENT FLAG DOCUMENT FLAG_DOCUMENT_10 -FLAG_DOCUMENT_12 -FLAG_DOCUMENT_14 -FLAG_DOCUMENT_16 -FLAG DOCUMENT 18 FLAG DOCUMENT 20 AMT REQ CREDIT BUREAU HOUR AMT REQ CREDIT BUREAU WEEK AMT_REQ_CREDIT_BUREAU_QRT

Analyse de fond



0.8

- 0.6

-0.2



Valeurs numériques Valeurs catégorielles



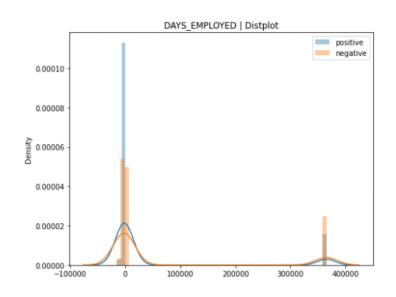
Corrélation valeurs numériques

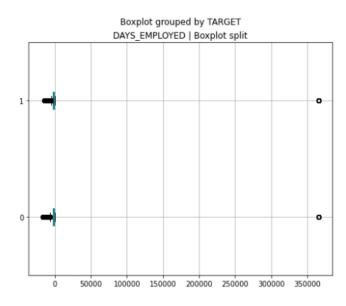


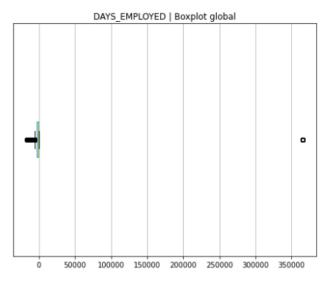
Traitement du jeu de données



Valeurs aberrantes

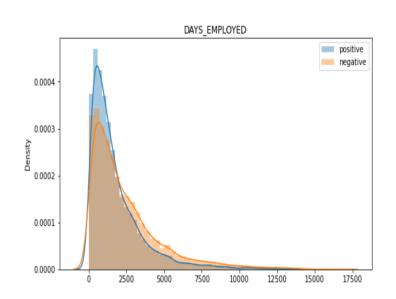


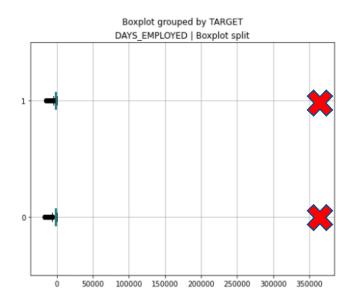


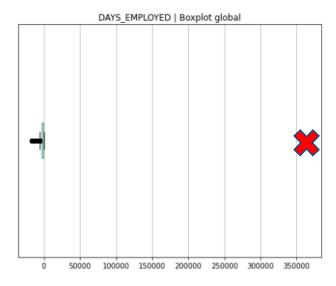




Valeurs aberrantes









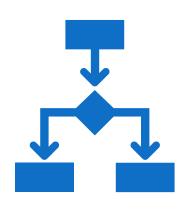
Imputation

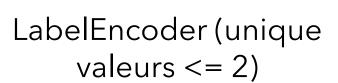




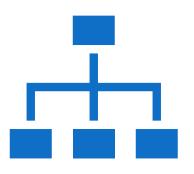
Reste à 0





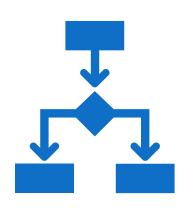


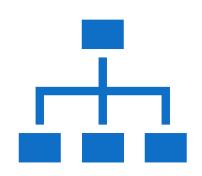
/	ANIMAL
1	CHIEN
2	CHAT



OneHotEncoder







LabelEncoder (unique valeurs <= 2)

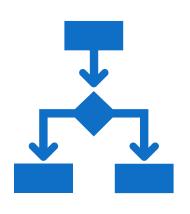
/	ANIMAL
1	0
2	1

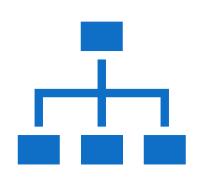
CHIEN = 0

• CHAT = 1









LabelEncoder (unique valeurs <= 2)

/	ANIMAL
1	0
2	1

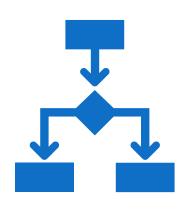
• CHIEN = 0

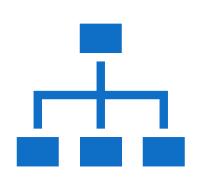
CHAT = 1

OneHotEncoder

	ANIMAL_
1	CHIEN
2	CHAT







LabelEncoder (unique valeurs <= 2)

/	ANIMAL
1	0
2	1

• CHIEN = 0

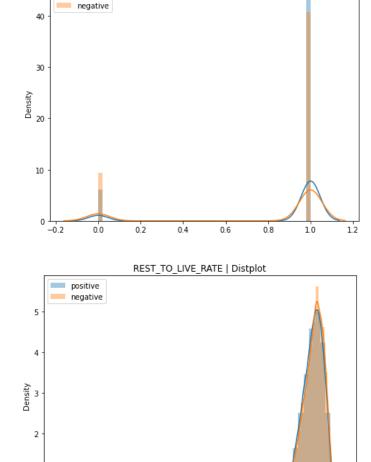
• CHAT = 1

OneHotEncoder

	ANIMAL_CHIEN	ANIMAL_CHAT
1	1	0
2	0	1



Création de nouvelles colonnes



-0.25

0.00

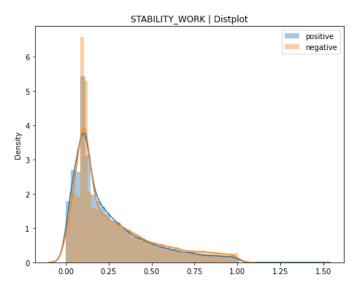
0.25

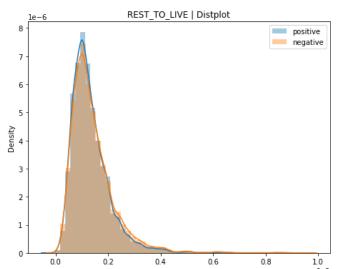
0.50

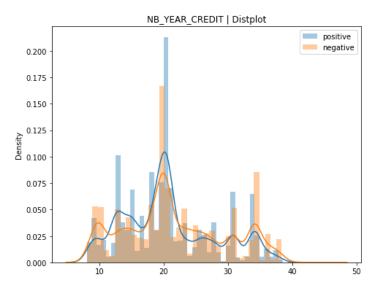
0.75

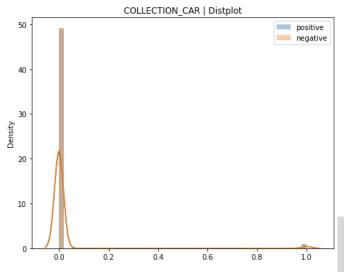
FLAG WORK | Distplot

positive











Sélections des données







BORUTA



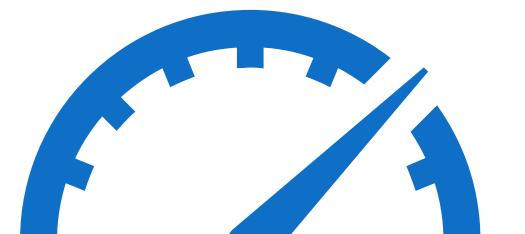
STATISTIQUE



Modélisation



Métrique



• F β -score (β à 2)

• Recall

• Precision



Choix des modèles

SupportVectorClassifier

LogisticRegression

KNeighborsClassifer

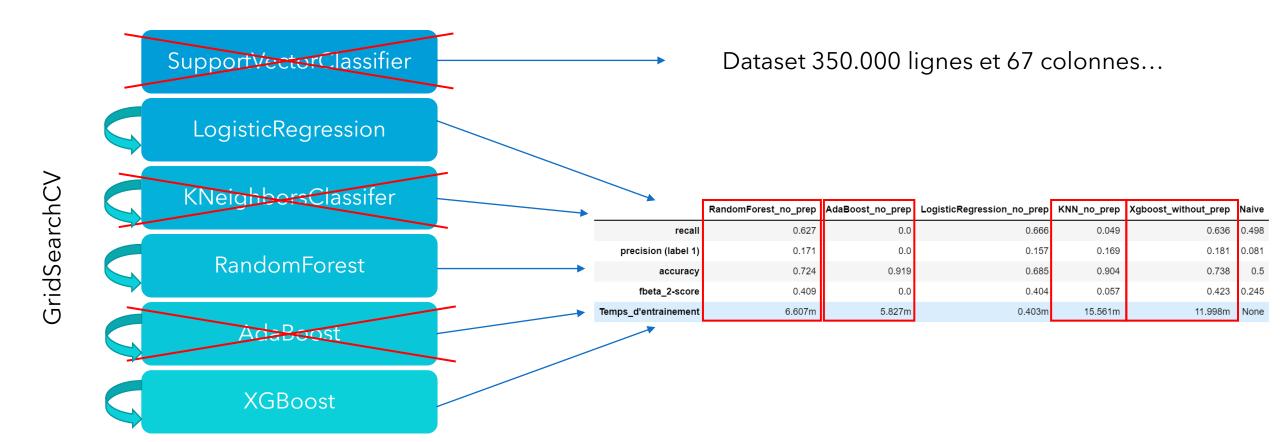
RandomForest

AdaBoost

XGBoost



Tests itératifs





Optimisation du modèle

RandomForest



• Hyperparamètres finaux :

➤N_estimators : 250

➤ Criterion : « entropy »

➤ Max_depth: 10

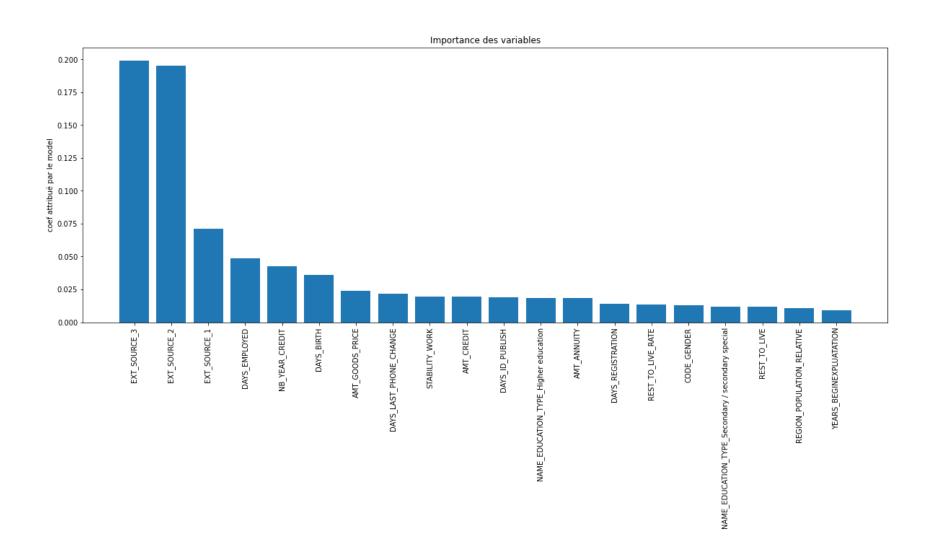
➤Class_weight : balanced



Explication du modèle



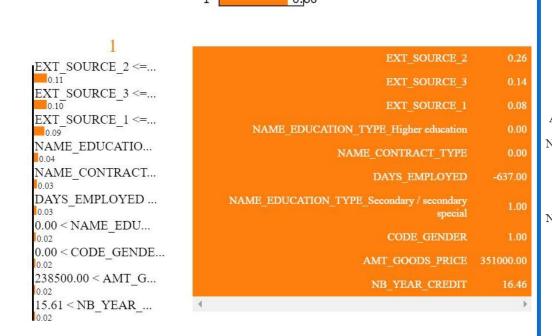
Explication simple

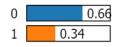


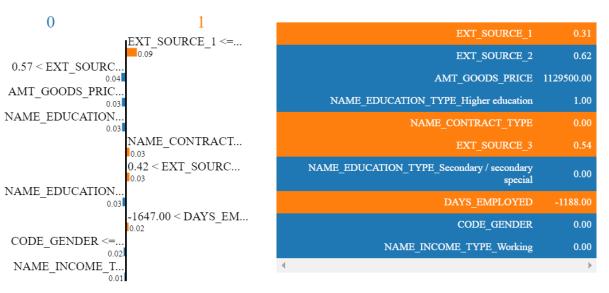


Explication par LIME

0.20

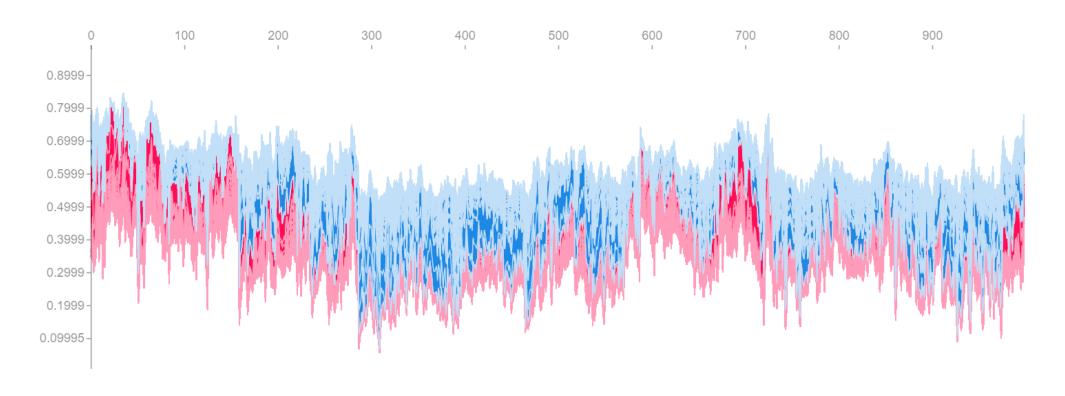




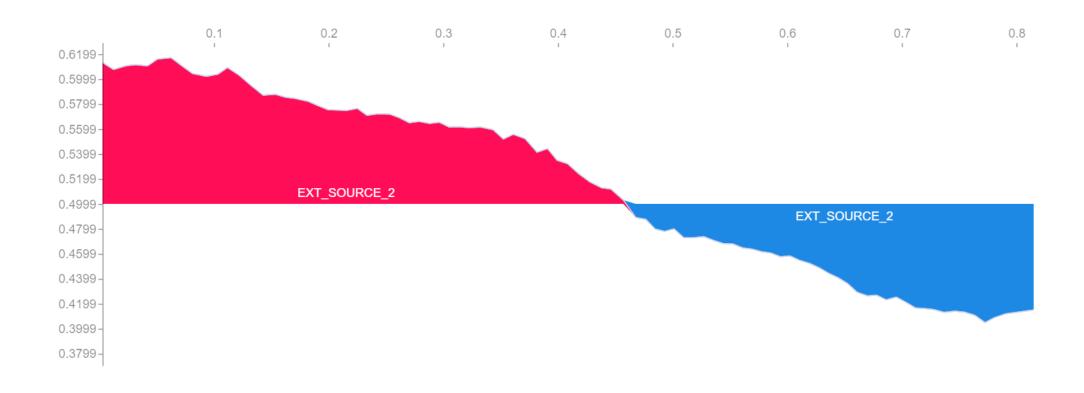




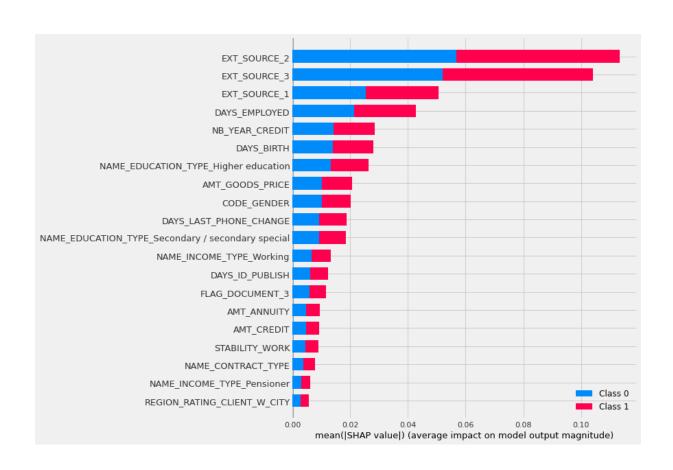
0



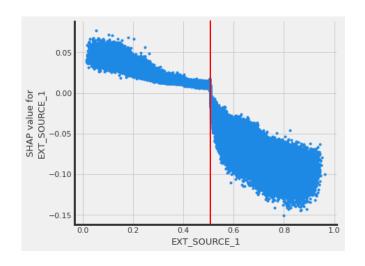


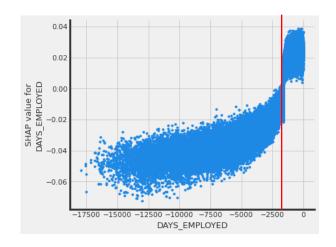


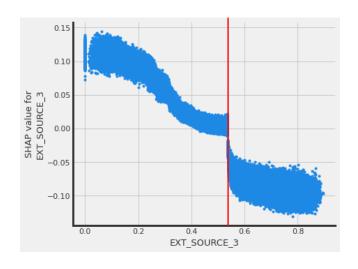


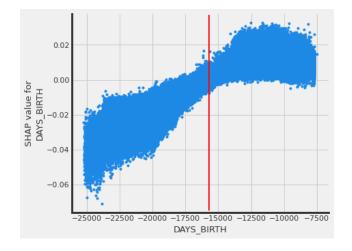


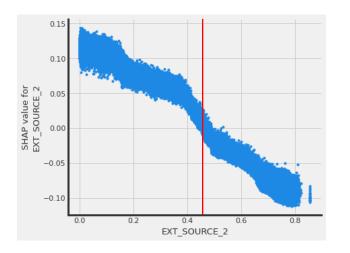














Conclusion



application_{train|test}.csv Main tables – our train and test samples Target (binary) Info about loan and loan applicant at application time · - · - · - · - SK_ID_CURR · --SK_ID_CURRbureau.csv previous application.csv Application data of client's Application data from previous previous loans in Home Credit loans that client got from other · Info about the previous loan institutions and that were parameters and client info at reported to Credit Bureau SK_ID_CURR time of previous application One row per client's loan in One row per previous Credit Bureau application -SK_ID_PREV-SK_ID_BUREAU SK_ID_PREV bureau balance.csv POS_CASH_balance.csv instalments_payments.csv credit_card_balance.csv · Past payment data for each · Monthly balance of · Monthly balance of · Monthly balance of installments of previous credits credits in Credit client's previous client's previous loans in Home Credit in Home Credit related to loans Bureau credit card loans in in our sample Behavioral data Behavioral data Home Credit Behavioral data Behavioral data

Conclusion



Score 0,408



Prendre tous les fichiers



Premier projet ML



Merci de votre attention, avez-vous des questions?

