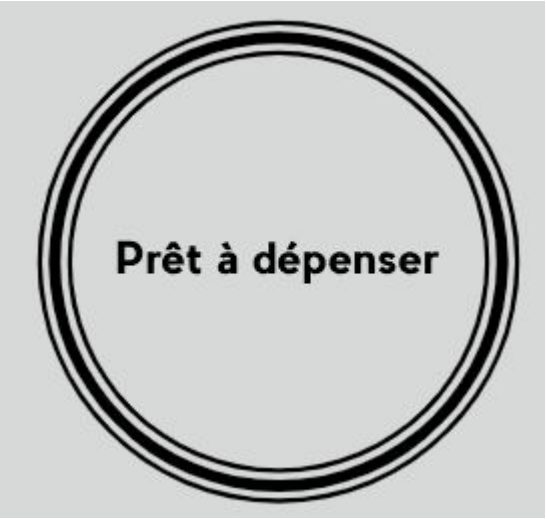
Ingénierie IA

Modèle de Scoring

Kindi BALDE

Présentation des données - 1/3

Problématique métier



"Prêt à dépenser", vous êtes une entreprise qui octroie des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt.

Pour accorder un crédit à la consommation, vous calculez la probabilité qu'un client le rembourse, ou non.

Vous souhaitez avoir à disposition un algorithme de scoring pour vos chargés de clients afin de les aider à décider si un prêt peut être accordé à un client.

Vous nous avez mis à dispositions plusieurs sources de données.

Notre objectif :
Livrer un modèle de scoring qui permet, en utilisant les informations sur un client de dire, à travers un score, si un client est fiable ou pas (bon ou mauvais)

Présentation des données - 2/3

Présentation du projet

Nous allons pour ce livrable, travailler sur le principal data set: ***application_train***

Sommaire

- **application_train/application_test**: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
- **bureau**: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
- **bureau_balance**: monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
- **previous_application**: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV.
- **POS_CASH_BALANCE**: monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
- **credit_card_balance**: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
- **installments_payment**: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

Présentation des données - 3/3

Présentation du projet

Sur la table *application_train*, chaque ligne est unique et représente un client.

Un client lambda, a donc sur sa ligne, des informations sur sa situation financière, son âge, son genre, son style de vie, son score, etc....

Le score prend 0 ou 1, 0 le client a remboursé son prêt, 1 le client n'a pas remboursé.

Sommaire

1. Garbage In Garbage Out

```
1 app_train_raw.head(3)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
0	100002	1	Cash loans	M	N	Y	0
1	100003	0	Cash loans	F	N	N	0
2	100004	0	Revolving loans	M	Y	Y	0

stats	
nbRows	307511
nbColumns	123
float64	65
int64	42
object	16

Nous avons nettoyé les données comme suit :

- Pour les variables qualitatives (catégorielles), contenant des manquants, nous avons créé des flag “Unknown”.
- Pour les variables de type numérique, nous avons faits des imputations par la moyenne.

Les données sont déséquilibrés. nous ferons du under sampling pour remédier à cela, vu que nous avons suffisamment de données pour faire l’entraînement.

```
1 app_train.TARGET.value_counts(normalize=True)*100
```

```
0    91.926961
1     8.073039
```

```
prep.missingValues(app_train_raw.select_dtypes('object'))
```

	Missing Values	% of Total Values
FONDKAPREMONT_MODE	210295	68.4
WALLSMATERIAL_MODE	156341	50.8
HOUSETYPE_MODE	154297	50.2
EMERGENCYSTATE_MODE	145755	47.4
OCCUPATION_TYPE	96391	31.3
NAME_TYPE_SUITE	1292	0.4

```
prep.missingValues(app_train_raw.select_dtypes(['int64', 'float64']))
```

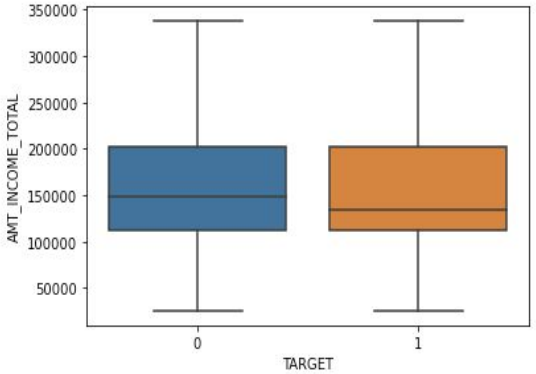
	Missing Values	% of Total Values
COMMONAREA_MODE	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MEDI	214865	69.9
NONLIVINGAPARTMENTS_AVG	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_MEDI	213514	69.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4

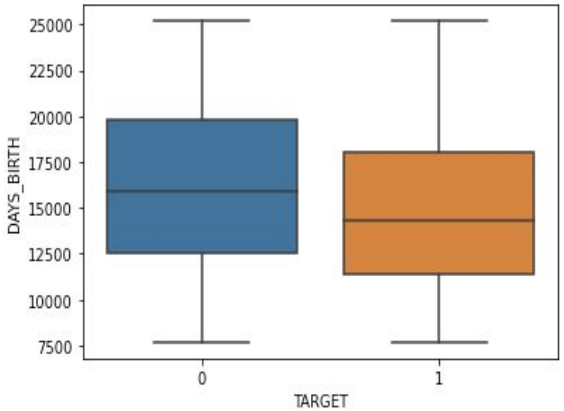
2. Analyse Exploratoire des données - 1/2

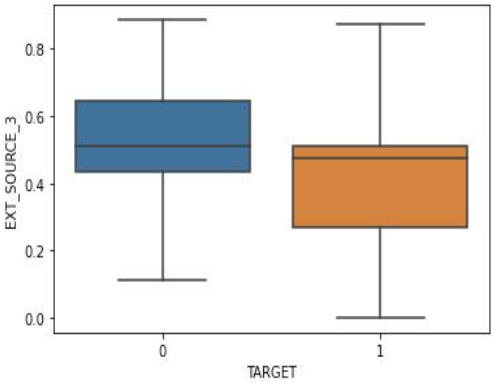
Quelques graphes

Nous voyons plus ou moins la pertinence de certaines variables. Les clients peuvent potentiellement être bien discriminer par ces variables.

On voit qu'à priori, l'âge, le niveau de revenue, sont entre autres potentiellement des variables importantes dans un processus prêt.



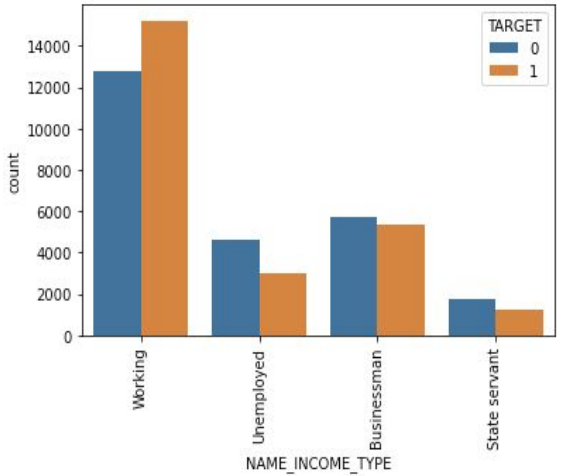


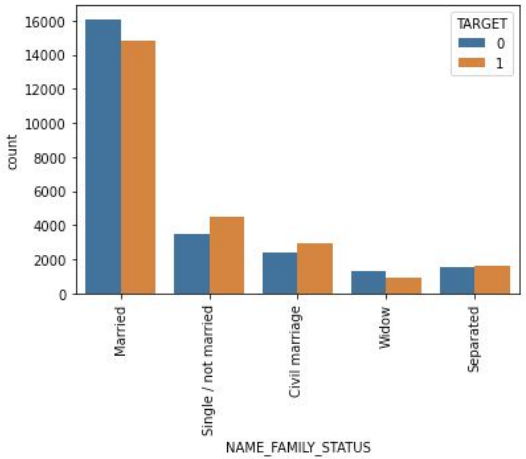


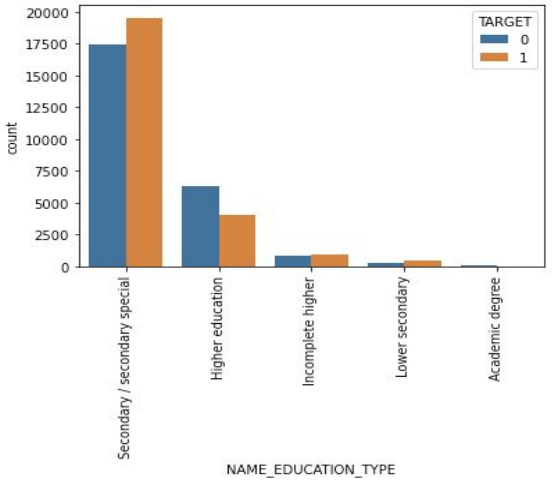
2. Analyse Exploratoire des données - 2/2

Quelques graphes

- Les Businessmen et les fonctionnaires ont tendance à bien gérer; les travailleurs ressortent comme étant des mauvais élèves; les chômeurs ont intérêt à être sage.
- Les personnes mariées et les veuves ont tendance à bien gérer; les célibataires ressortent comme étant des personnes potentiellement à risques.
- Les personnes à niveau d'éducation bas ont tendance à ressortir comme étant de mauvais payeurs.







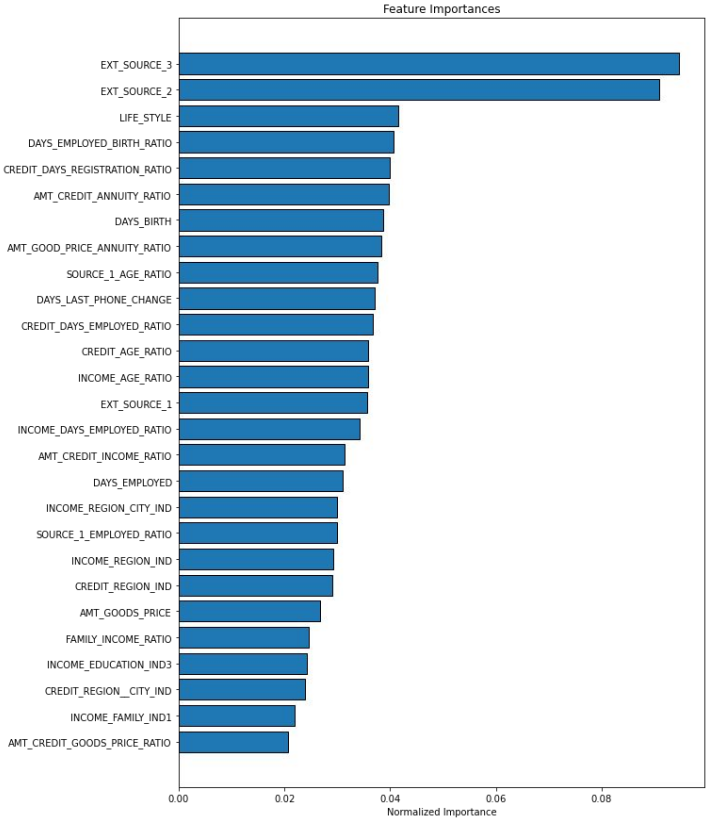
3. Sélection de variables

Notre sélection de variables a été faite sur trois axes :

- Suite à une analyse de corrélation il est ressorti que les variibles suivantes sont très pertinentes : DAYS_BIRTH, DAYS_EMPLOYED, AMT_GOODS_PRICE, DAYS_LAST_PHONE_CHANGE, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3 .

- Au cours de cette même analyse, il est ressorti que beaucoup sont corrélées. Ce sont 36 variables quantitatives qui décrivent le style de vie du client (son type d'appartement, le standing de son immeuble, etc..). Une Analyse en Composante principale, confirme cela en regroupant ces variables sur un axe. Une variable nommée "LIFE_STYLE" a été créé avec ce process.

- La construction de nouvelles variables en se basant sur la connaissance métier : INCOME_FAMILY_IND1, INCOME_EDUCATION_IND3, INCOME_REGION_IND, CREDIT_REGION_IND, INCOME_REGION_CITY_IND, AMT_GOOD_PRICE_ANNUITY_RATIO, AMT_CREDIT_ANNUITY_RATIO, AMT_CREDIT_GOODS_PRICE_RATIO, AMT_CREDIT_INCOME_RATIO, INCOME_AGE_RATIO, FAMILY_INCOME_RATIO, CREDIT_AGE_RATIO, CREDIT_DAYS_REGISTRATION_RATIO, CREDIT_DAYS_EMPLOYED_RATIO, INCOME_DAYS_EMPLOYED_RATIO, DAYS_EMPLOYED_BIRTH_RATIO, SOURCE_1_AGE_RATIO, SOURCE_1_EMPLOYED_RATIO



3. Sélection de variables

Définition de la construction de certaines variables

INCOME_FAMILY_IND1 AMT_INCOME_TOTAL*'NAME_FAMILY_STATUS_Married/CNT_FAM_MEMBERS

INCOME_EDUCATION_IND3 MT_INCOME_TOTAL* NAME_EDUCATION_TYPE_Secondary

INCOME_REGION_IND AMT_INCOME_TOTAL * REGION_RATING_CLIENT / AMT_GOODS_PRICE

CREDIT_REGION_IND AMT_CREDIT * REGION_RATING_CLIEN T/ AMT_GOODS_PRICE

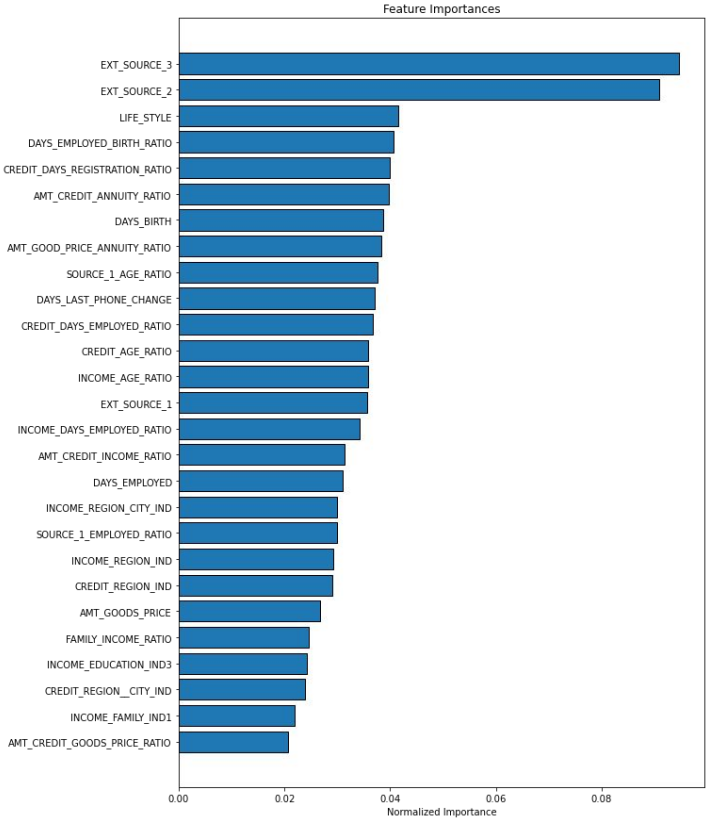
INCOME_REGION_CITY_IND AMT_CREDIT * REGION_RATING_CLIENT / AMT_GOODS_PRICE

AMT_GOOD_PRICE_ANNUITY_RATIO AMT_INCOME_TOTAL*REGION_RATING_CLIENT_W_CITY / AMT_GOODS_PRICE

AMT_CREDIT_ANNUITY_RATIO AMT_GOODS_PRICE / AMT_ANNUITY

AMT_CREDIT_GOODS_PRICE_RATIO AMT_CREDIT / AMT_GOODS_PRICE

AMT_CREDIT_INCOME_RATIO AMT_CREDIT / AMT_INCOME_TOTAL



4. Sélection de modèles de classification

Plusieurs modèles de classifications

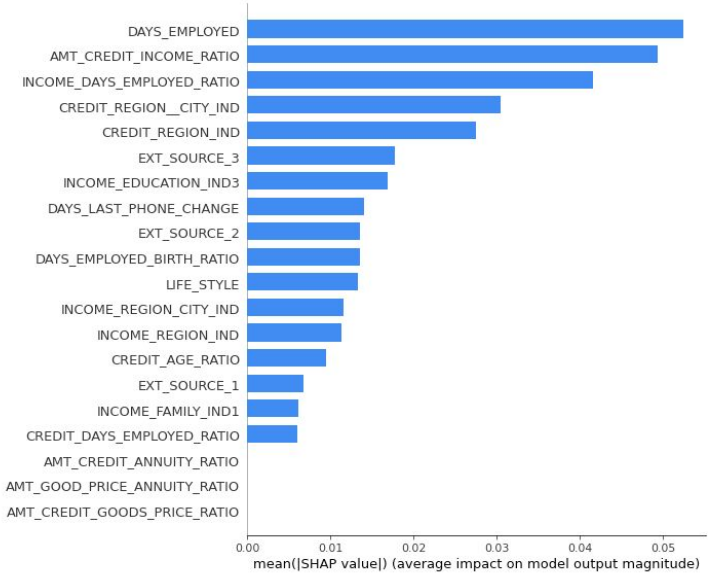
Nous avons testés cinq modèles de classification. Pour l'optimisation de chaque modèle, nous avons utilisé la méthode d'optimisation de Bayes, qui consiste à utiliser une certaine information à priori sur les performances des paramètres de chaque modèle pour trouver les meilleurs. Elle est rapide et efficace.

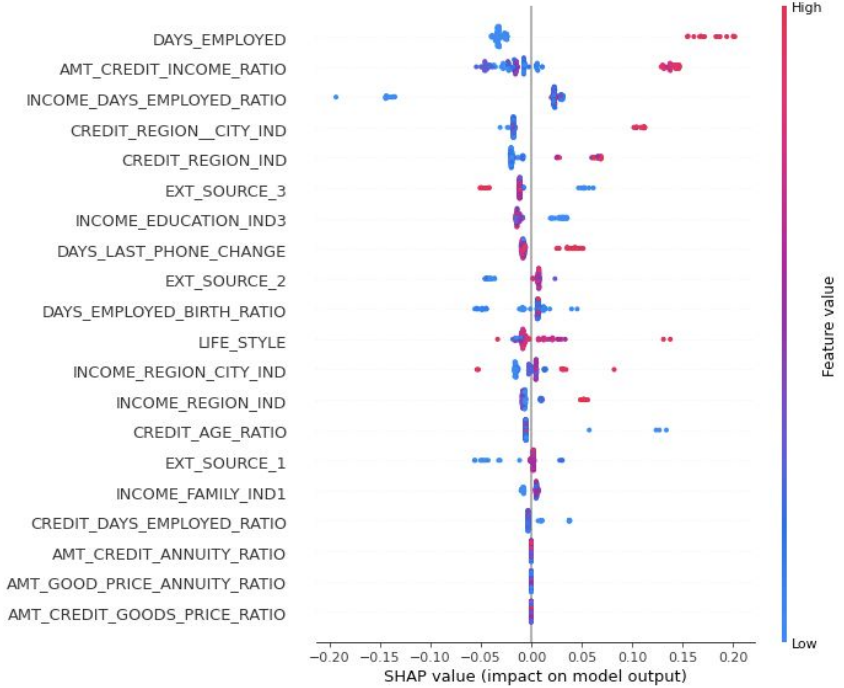
MODEL	PERF_TRAINING_SET	MODEL_SCORE_TEST_SET	PRECISION_SCORE_TEST_SET	RECALL_SCORE_TEST_SET	F1_SCORE_TEST_SET
LR	67.933263	66.988099	67.452653	65.649414	66.538818
KNN	65.771832	65.688129	65.636408	65.844727	65.740402
SVM	68.026454	67.671651	68.082448	66.528320	67.296413
RF	68.687810	68.141593	68.427579	67.358398	67.888780
XGB	69.165790	68.654257	69.080919	67.529297	68.296296

5. Interprétation du modèle

nous avons utilisé la librairie Shap. Elle permet d'identifier et de mettre en valeur la contribution de chaque variable explicative dans la construction du modèle. Cela donne ainsi un proxy d'interprétabilité du modèle.

Variables sont ordonnées par ordre d'importance décroissante. Et notamment sur l'image de droite, nous par exemple que la variable DAYS_EMPLOYED est très importante dans la séparation des clients et que ses grandes valeurs contribuent significativement à la qualité de la prévision.





5. Interprétation du modèle

Présentation des données

Toujours avec la librairie Shap, nous avons construit des “waterfall”.

- Pour le Score 1, pour ceux qui ne remboursent pas, on voit que certaines variables permettent de mieux les identifier. le CREDIT_AGE_RATIO (l’âge par rapport au montant du crédit) par exemple montre plus ce ratio est grand, moins on aura tendance à rembourser.
- Pour le Score 0, ci-dessous, on voit par exemple le lieu de résidence du client à un effet positif et permet de bien faire ressortir les bons payeurs.