

# **Implémentez un modèle de scoring**

**Prêt à Dépenser**





# Plan de Présentation

Axes principaux

- ❑ Contexte
- ❑ Data
- ❑ Nettoyage
- ❑ Modélisation
- ❑ API & Dashboard
- ❑ Axes d'amélioration



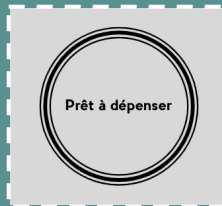
# CONTEXTE



# Le besoin et la ressource

## Besoin

- ❑ Michaël - Prêt à dépenser
- ❑ Modèle de Scoring
- ❑ Classification
- ❑ Transparence
- ❑ Dashboard



## Ressources

- ❑ PàD Dataset
- ❑ Kernel Kaggle
- ❑ Connaissances Machine Learning
- ❑ Volonté d'apprendre

The background is a solid teal color. In the top-left corner, there are three vertical bars of varying heights, each composed of three overlapping circles. In the bottom-right corner, there are four vertical bars of increasing height, each composed of three overlapping circles.

**DATA**

# Data

## Data Explorer

2.68 GB

- HomeCredit\_columns\_descr...
- POS\_CASH\_balance.csv
- application\_test.csv
- application\_train.csv
- bureau.csv
- bureau\_balance.csv
- credit\_card\_balance.csv
- installments\_payments.csv
- previous\_application.csv
- sample\_submission.csv

1

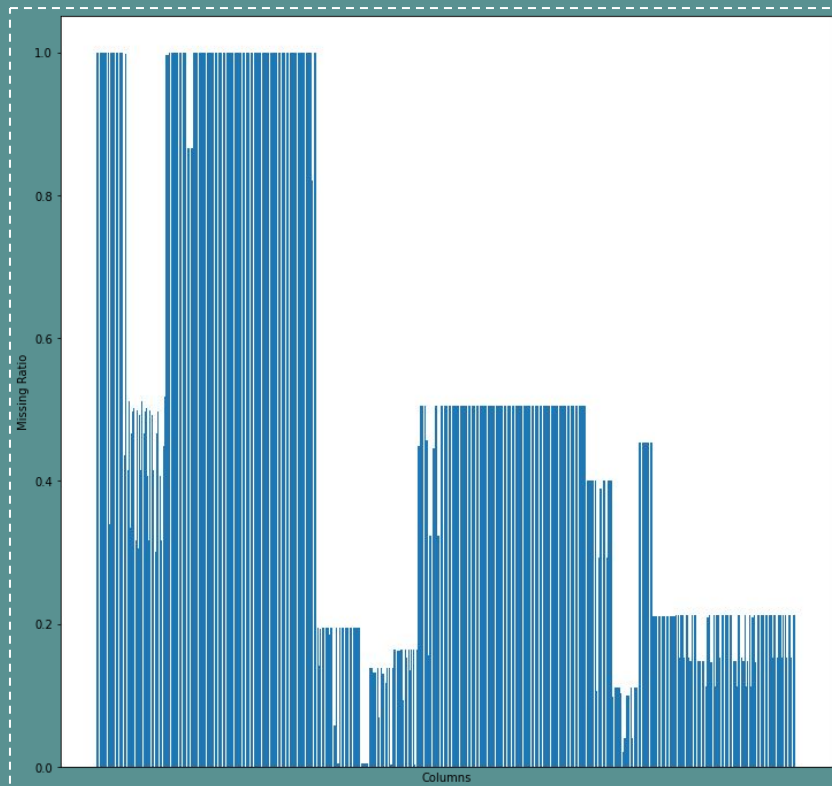
307507

id	name	type	value	date	status	...
1	POS_CASH_balance	float	0.0	2018-05-01	active	...
2	application_test	float	0.0	2018-05-01	active	...
3	application_train	float	0.0	2018-05-01	active	...
4	bureau	float	0.0	2018-05-01	active	...
5	bureau_balance	float	0.0	2018-05-01	active	...
6	credit_card_balance	float	0.0	2018-05-01	active	...
7	installments_payments	float	0.0	2018-05-01	active	...
8	previous_application	float	0.0	2018-05-01	active	...
9	sample_submission	float	0.0	2018-05-01	active	...

Noyau Kaggle

# Après l'utilisation du noyau Kaggle

- Données numériques
- Dummies sur les variables catégorielles
- Agrégations classiques (mean, max,...)



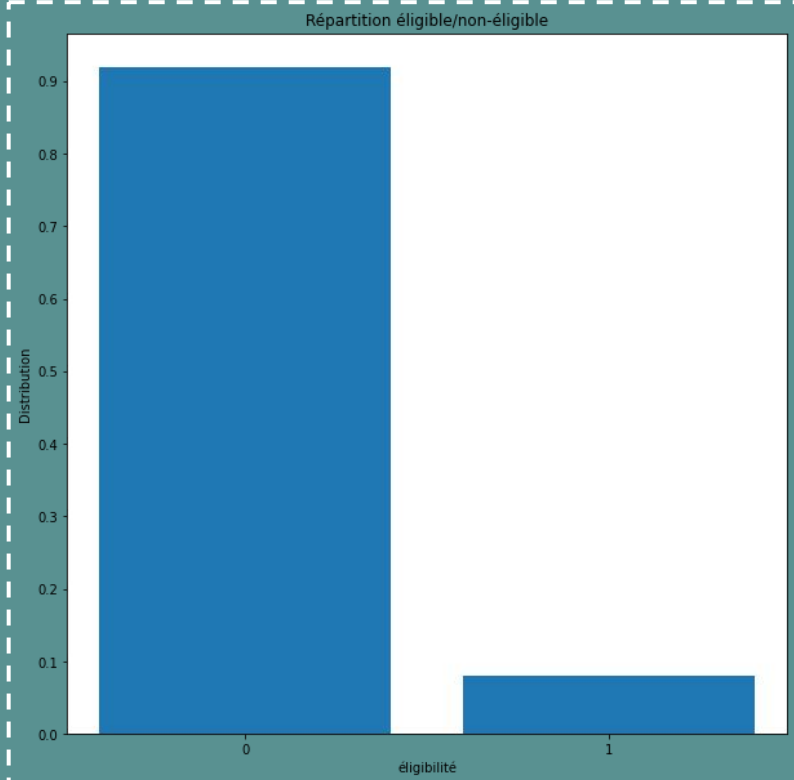
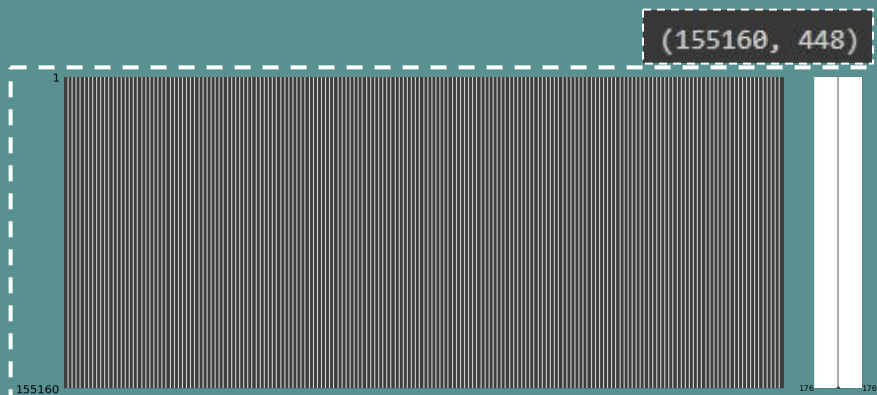


**NETTOYAGE**



# EDA et nettoyage

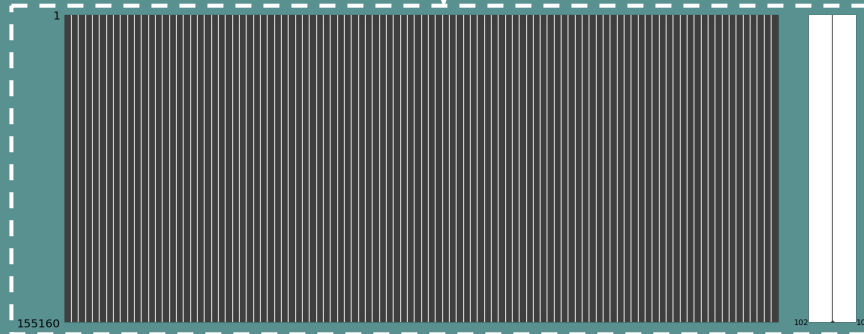
- Suppression des valeurs problématiques (outliers, -inf, etc...)
- Lignes / colonnes %manquant > 40%
- Remplissage de valeurs manquantes
  - catégorielles = mode
  - numériques = median



# Modélisation

## Feature Selection

	Feature	Pearson	Chi-2	RFE	Random Forest	LightGBM	Total
1	EXT_SOURCE_3	True	True	True	True	True	5
2	EXT_SOURCE_2	True	True	True	True	True	5
3	EXT_SOURCE_1	True	True	True	True	True	5
4	DEF_30_CNT_SOCIAL_CIRCLE	True	True	True	True	True	5
5	DAYS_ID_PUBLISH	True	True	True	True	True	5
...	...	...	...	...	...	...	...
96	APPROVED_HOUR_APPR_PROCESS_START_MAX	False	False	True	True	True	3
97	APPROVED_DAYS_DECISION_MIN	True	False	False	True	True	3
98	APPROVED_DAYS_DECISION_MEAN	True	False	False	True	True	3
99	APPROVED_DAYS_DECISION_MAX	True	False	False	True	True	3
100	APPROVED_AMT_CREDIT_MAX	False	False	True	True	True	3



(155160, 448)

```
1 df_model.shape  
(155160, 102)
```

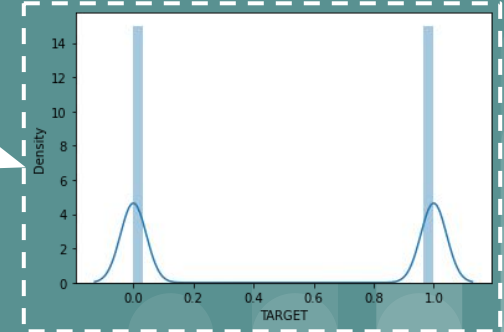
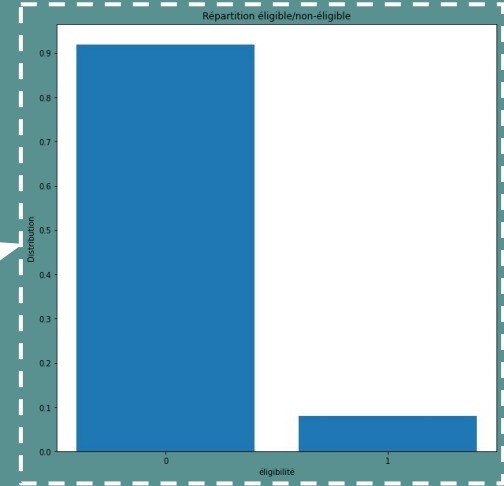
- Feature Selection
- Target Balancing

```
[ ] 1 y_train.value_counts()
```

```
0    106784  
1     9586  
Name: TARGET, dtype: int64
```

```
[ ] 1 y_train_res.value_counts()
```

```
1    106784  
0    106784  
Name: TARGET, dtype: int64
```



Keywords: SMOTENC, N-nearest-neighbours

## Fonction Coût

```
[ ] 1 def score_it(y, y_pred):  
    2     tn, fp, fn, tp = confusion_matrix(y, y_pred).ravel()  
    3     return (tn+tp +5*fp + 10*fn)/(tn+tp+fn+fp)
```

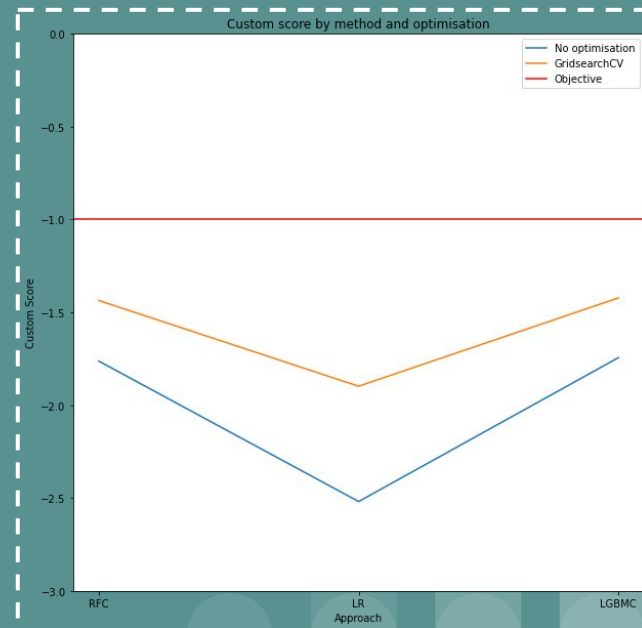
- Feature Selection
- Target Balancing
- Model Selection

	Method	RFC	LR	LGBMC
0	No Optimisation	-1.76298	-2.519026	-1.744161

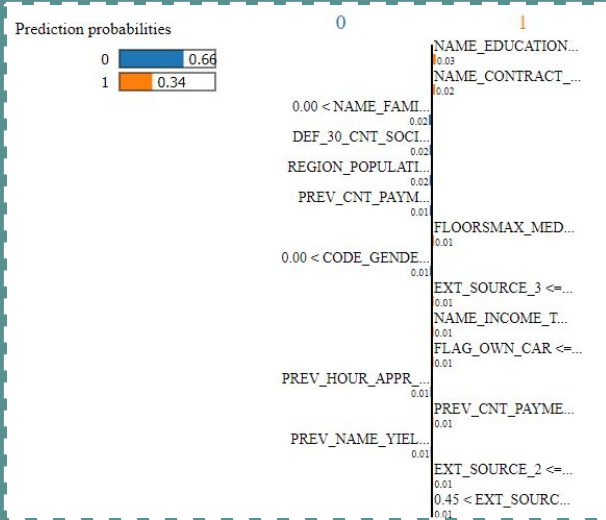
	RFC	LR	LGBMC
0	-1.436521	-1.898089	-1.422493

```
[ ] 1 df_result_pipe
```

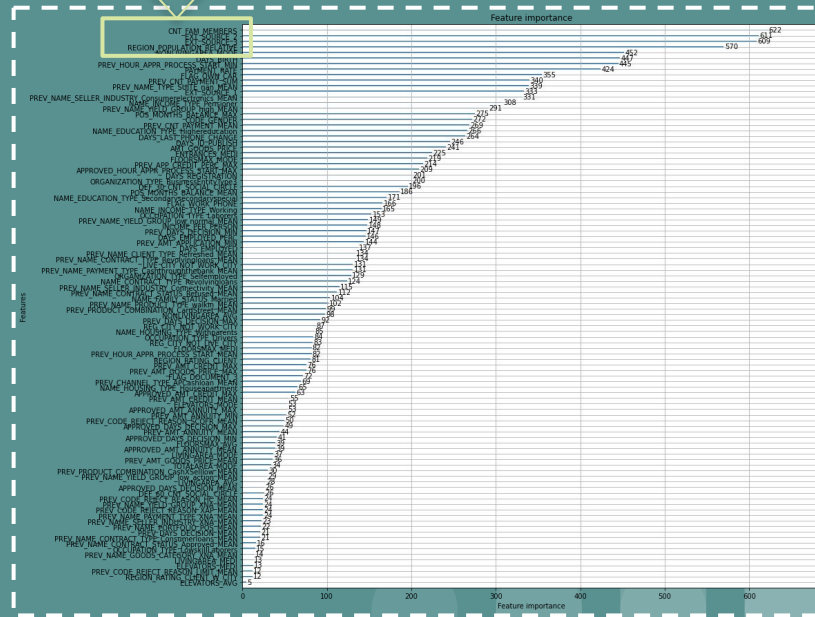
	model	Accuracy	Precision	Recall	F1	ROC_AUC	custom_score	best_param
0	LGBMClassifier(random_state=0)	0.951144	0.993149	0.909172	0.941325	0.966175	-1.422493	{'boosting_type': 'dart', 'learning_rate': 0.1...



- Feature Selection
- Target Balancing
- Model Selection
- Interprétabilité



CNT\_FAM\_MEMBERS  
EXT\_SOURCE\_2  
EXT\_SOURCE\_3  
REGION\_POPULATION  
NONLIVINGAREA\_MODE

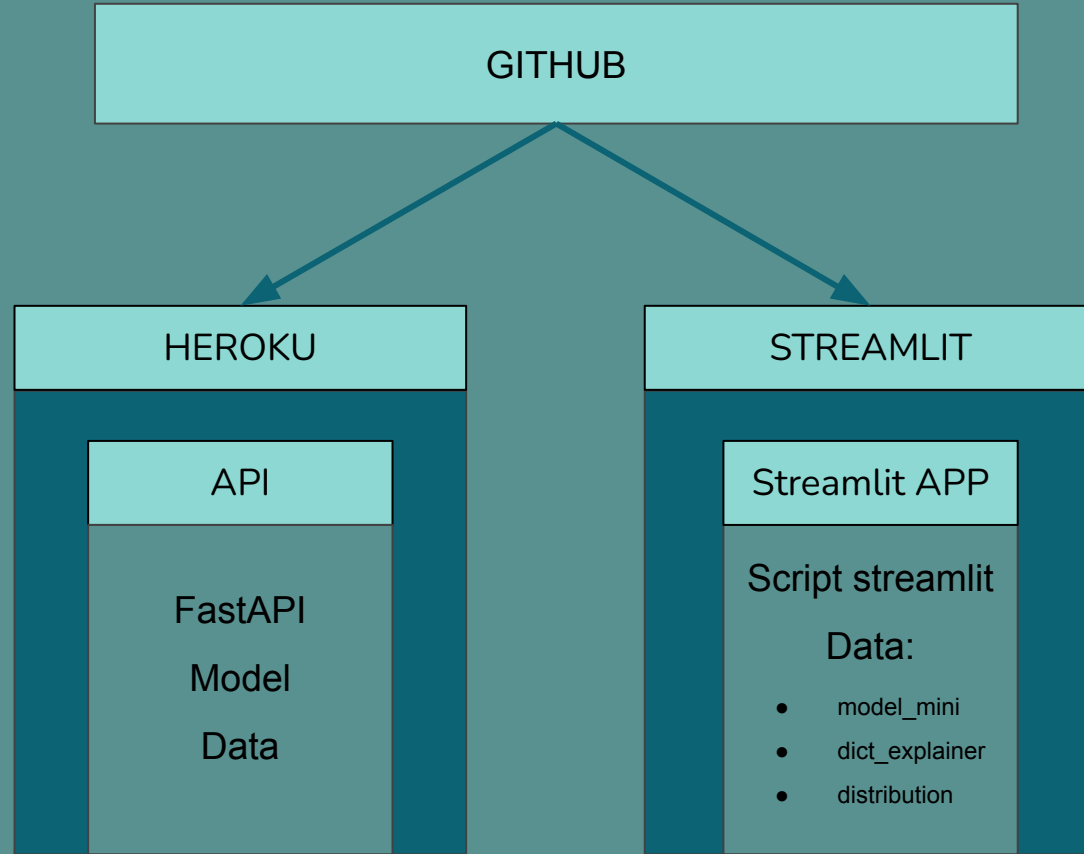




# **API & DASHBOARD**



□ Structure





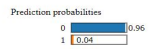
# ALLONS VOIR À QUOI ILS RESSEMBLENT

## DASHBOARD

### Client knowledge Dashboard

	EXT_SOURCE_3	EXT_SOURCE_2	EXT_SOURCE_1	DEF_30_CNT_SOCIAL_CIRCLE	DAYS_ID_PUBLI
0	0.5245	0.6504	0.5073	0.0000	

Eligible à: 0.96%



0

EXT\_SOURCE\_2 > 0.62

PREV\_CNT\_PAYM...

DEF\_30\_CNT\_SOC...

REGION\_POPULAT...

PREV\_NAME\_TYPE...

PREV\_NAME\_YIEL...

1

NAME\_CONTRACT...

NAME\_EDUCATION...

NAME\_FAMILY\_STA...

FLAG\_OWN\_CAR <=...

FLOORSMAX\_MED...

#### Feature

NAME\_CONTRACT\_TYPE\_Revolvingloans

NAME\_EDUCATION\_TYPE\_Highereducation

EXT\_SOURCE\_2

NAME\_FAMILY\_STATUS\_Married

PREV\_CNT\_PAYMENT\_SUM

DEF\_30\_CNT\_SOCIAL\_CIRCLE

FLAG\_OWN\_CAR

FLOORSMAX\_MEDI

REGION\_POPULATION\_RELATIVE

PREV\_NAME\_TYPE\_SUITE\_sas\_MEAN

PREV\_NAME\_YIELD\_GROUP\_high\_MEAN

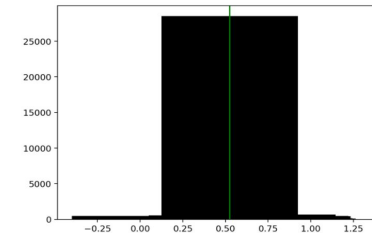
PREV\_CNT\_PAYMENT\_MEAN

CODE\_GENDER

Client	
100006	
CNT_FAM_MEMBERS	2.0000
EXT_SOURCE_2	0.6504
NONLIVINGAREA_MODE	0.0011
EXT_SOURCE_3	0.5245
REGION_POPULATION_RELATI...	0.0080

### DATAVIZ

Variable  
EXT\_SOURCE\_3



## API

### FastAPI 0.1.0 OAS3

/openapi.json

#### default

GET

/ Index

GET

/ {name} Get Name

POST

/predict Predict

POST

/predictProba0 Predictproba0

POST

/predictProba1 Predictproba1

# AXES D'AMÉLIORATION

- ❑ Cloud payant ( optimisation temps de chargement + volume data)
- ❑ Plus d'échange avec l'équipe (Produit plus adapté, feature plus mises en valeur)
- ❑ Meilleure qualité Dashboard: à voir avec les designers de l'équipe, descriptions de features, meilleure mise en évidence éligible/non-éligible.
- ❑ BackEnd Dashboard plus optimisé (Utilisation des caches Streamlit, graphes, etc)