Implémentez un modèle de scoring

Prêt à Dépenser

Plan de Présentation

Axes principaux

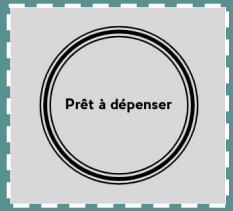
- ❏ Contexte
- ❏ Data
- ❏ Nettoyage
- ❏ Modélisation
- ❏ API & Dashboard
- ❏ Axes d’amélioration

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

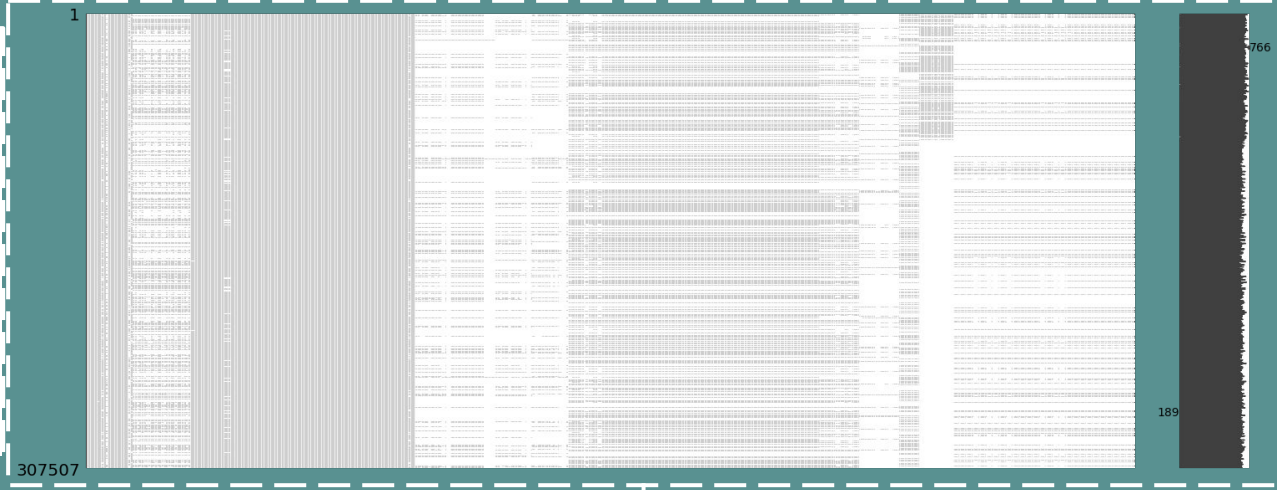
DATA

Data

Data Explorer

2.68 GB

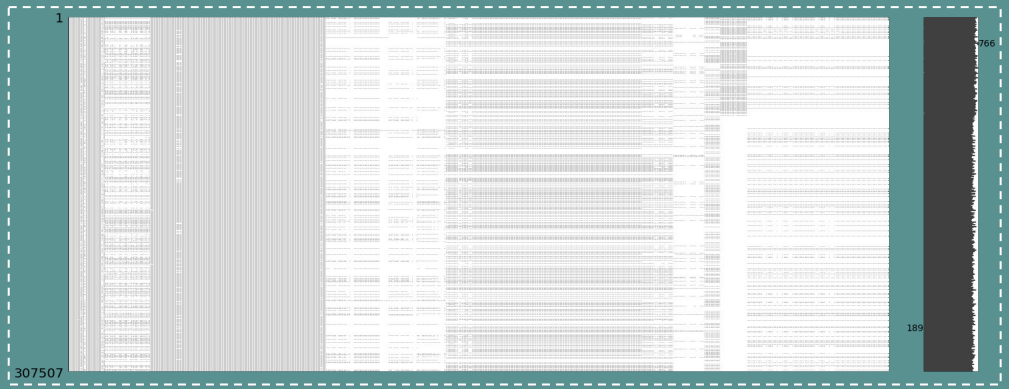
- HomeCredit_columns_descr...
- POS_CASH_balance.csv
- application_test.csv
- application_train.csv
- bureau.csv
- bureau_balance.csv
- credit_card_balance.csv
- installments_payments.csv
- previous_application.csv
- sample_submission.csv

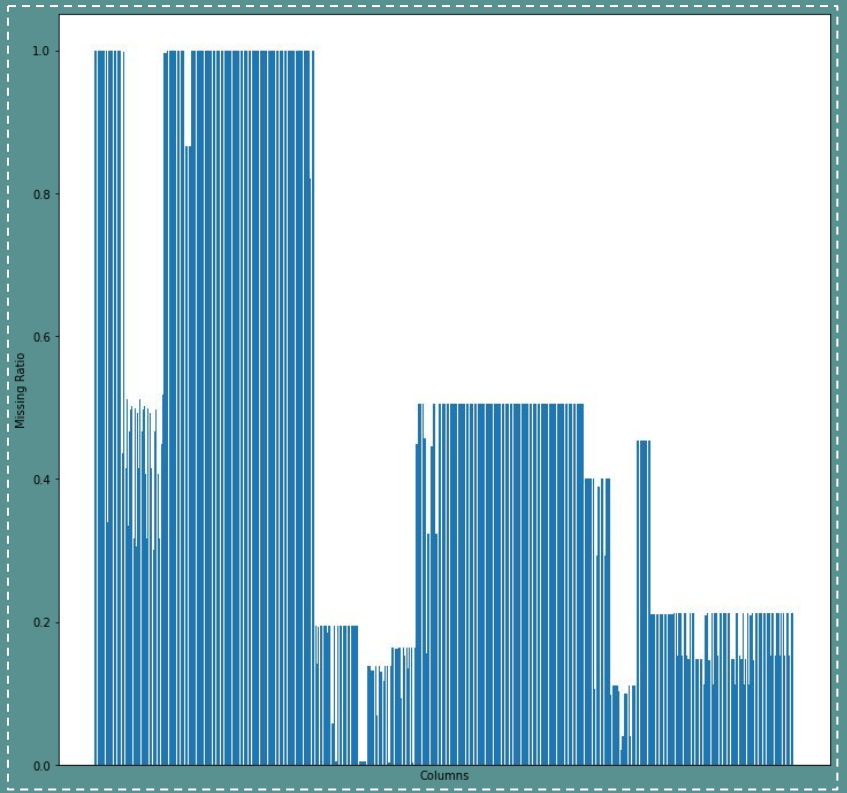


Noyau Kaggle

Après l'utilisation du noyau Kaggle

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

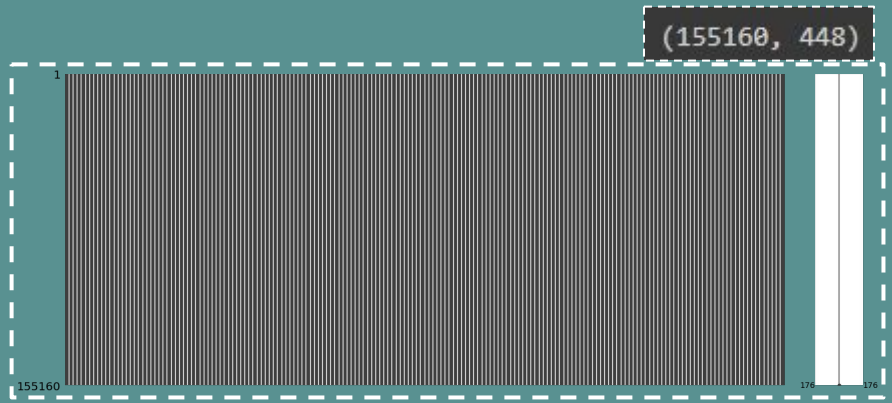


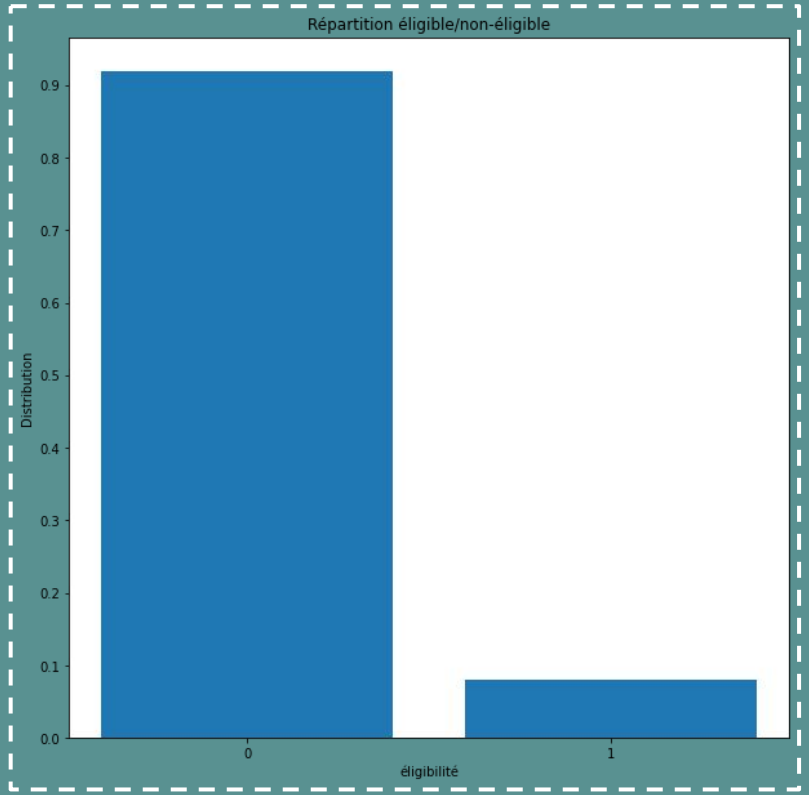


NETTOYAGE

EDA et nettoyage

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

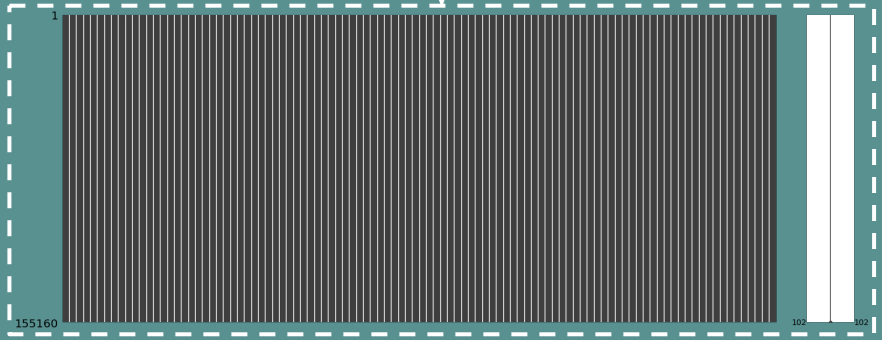




Modélisation

- 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)
```

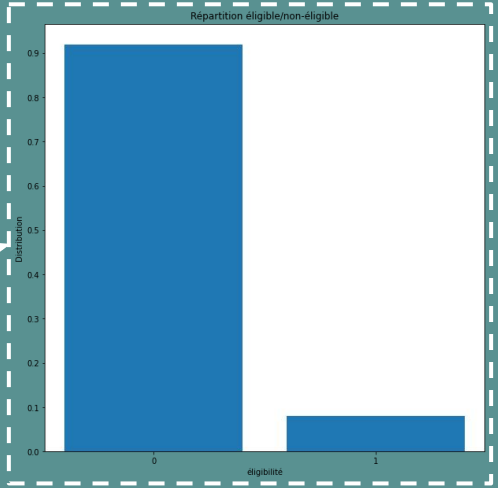
- Feature Selection
- Target Balancing

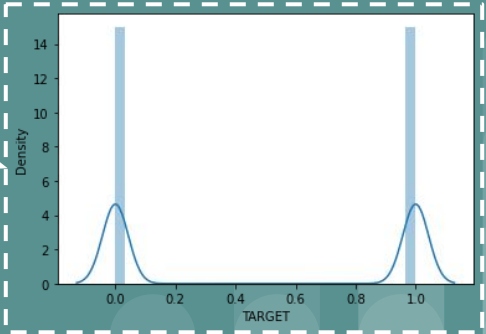
```
[ ]  1 y_train.value_counts()

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)
```

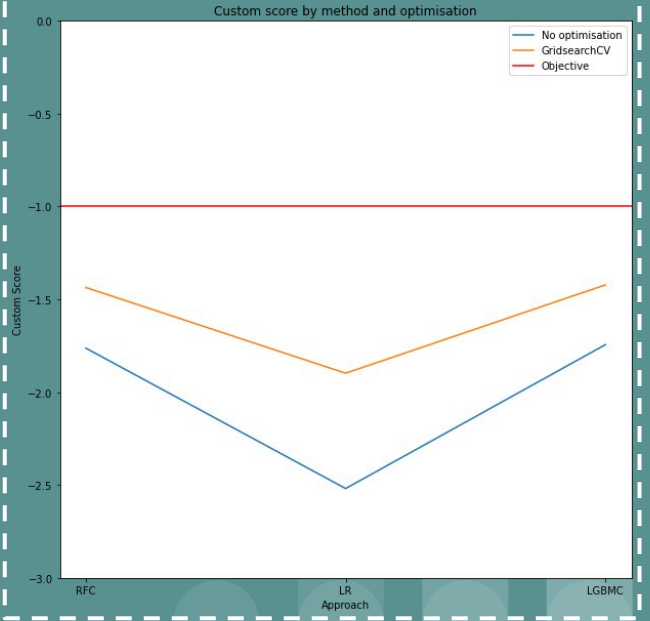
- Feature Selection
- Target Balancing
- Model Selection

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

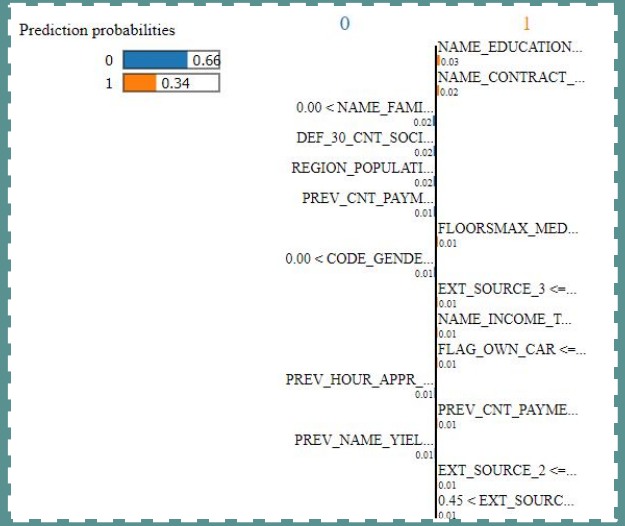
	RFC	LR	LGBMC
0	-1.436521	-1.898089	-1.422493

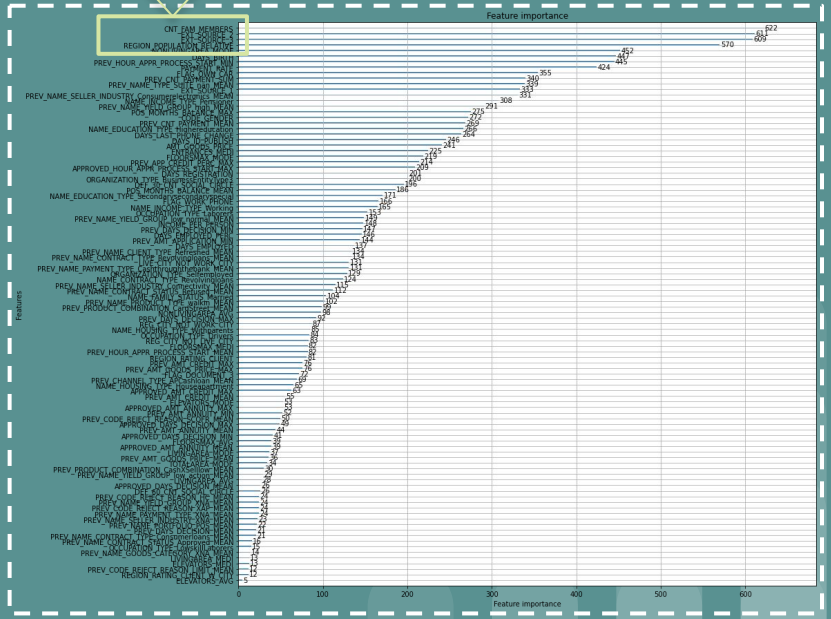
```
[ ]  1 df_result_pipe
```

	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...



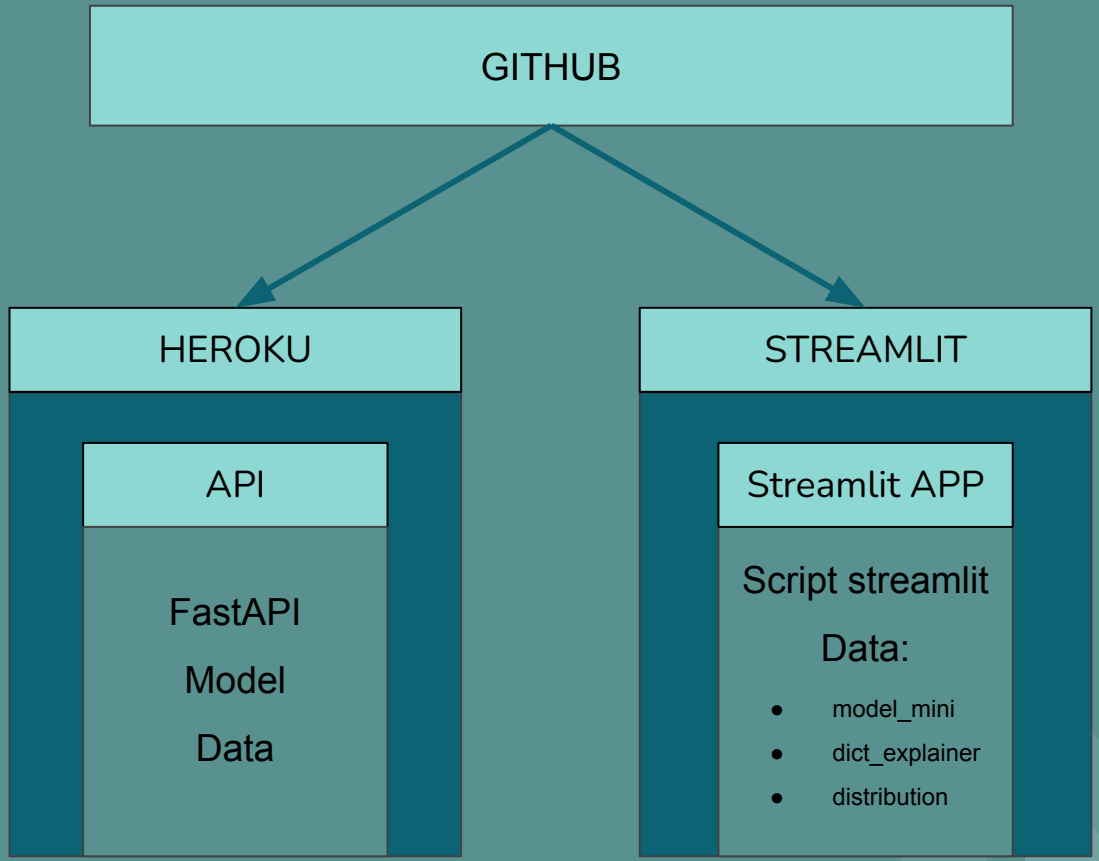
- ❏ Feature Selection
- ❏ Target Balancing
- ❏ Model Selection
- ❏ Interprétabilité





API & DASHBOARD

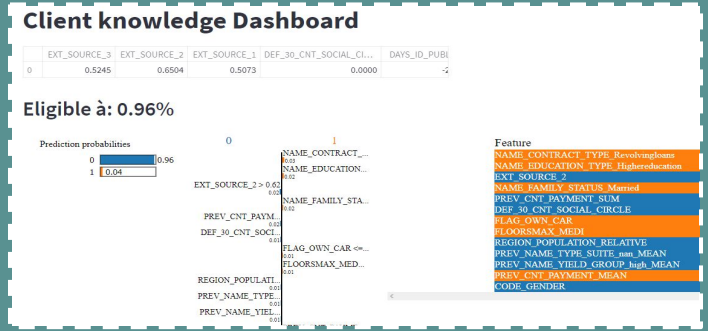
❑ Structure

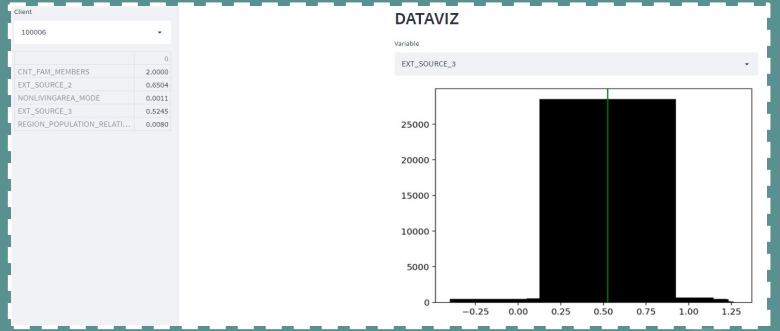

GITHUB
HEROKU
API
FastAPI
Model
Data
STREAMLIT
Streamlit APP
Script streamlit
Data:
model_mini
dict_explainer
distribution

ALLONS VOIR À QUOI ILS RESSEMBLENT

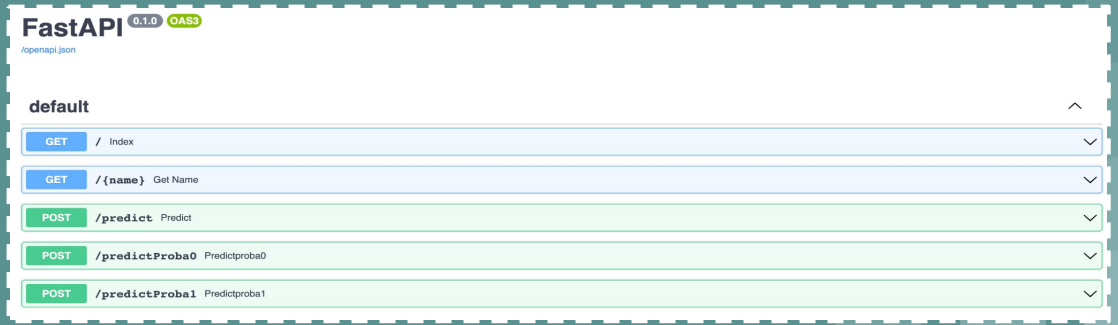
- Structure
- API &
  Dashboard

DASHBOARD





API



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)