

OPENCCLASSROOMS



pôle emploi

Projet 7.

Implémentez un modèle de scoring

Dashboard **PRET A DEPENSER**

Please select a Client in the sidebar in the left



Sommaire:

1. *Mission*
2. *Données*
3. *Metrics*
4. *Observation*
5. *Preparation*
6. *Modèles*
7. *Résultats*
8. *API & DashBoard*
9. *Conclusion*

Implémentez un modèle de scoring

Presentation Mission

Dashboard **PRÊT A DÉPENSER**

Please select a Client in the sidebar in the left



L'entreprise souhaite **mettre en œuvre un outil de “scoring crédit” pour calculer la probabilité** qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Elle souhaite donc développer un **algorithme de classification** en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de **transparence** vis-à-vis des décisions d'octroi de crédit. Cette demande de transparence des clients va tout à fait dans le sens des valeurs que l'entreprise veut incarner.

Prêt à dépenser décide donc de **développer un dashboard interactif** pour que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

Presentation Mission

Dashboard **PRET A DEPENSER**

Please select a Client in the sidebar in the left

Prêt à dépenser



Votre mission

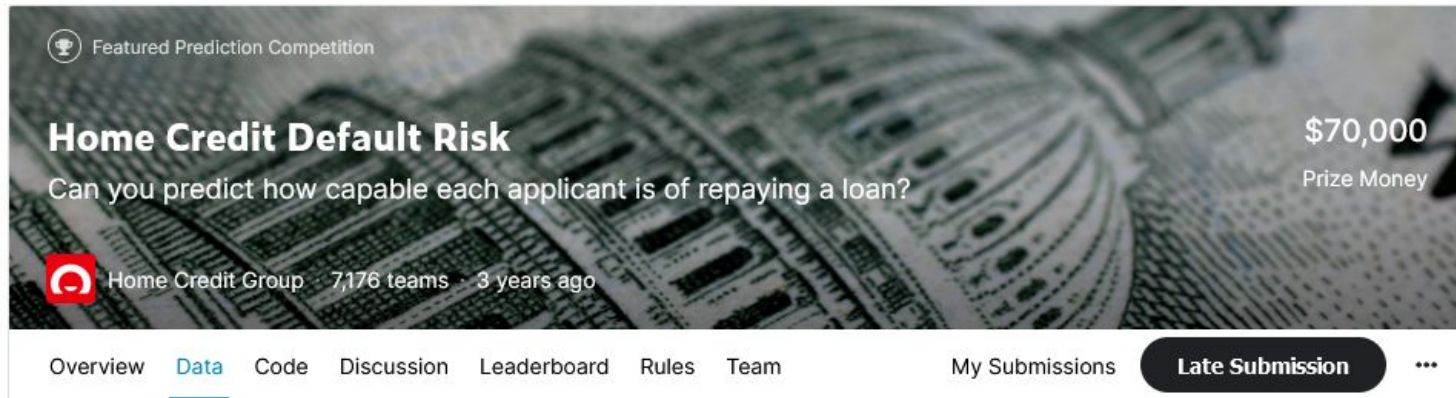
1. Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.
2. Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client des chargés de relation client.

Michaël, votre manager, vous incite à sélectionner un kernel Kaggle pour vous faciliter la préparation des données nécessaires à l'élaboration du modèle de scoring. Vous analyserez ce kernel et l'adapterez pour vous assurer qu'il répond aux besoins de votre mission.



Vous pourrez ainsi vous focaliser sur l'élaboration du modèle, son optimisation et sa compréhension.

Les données Implémentez un modèle de scoring





The screenshot shows the top section of a Kaggle competition page. At the top left, it says 'Featured Prediction Competition' with a star icon. The main title is 'Home Credit Default Risk' in large bold letters. Below it, the question is 'Can you predict how capable each applicant is of repaying a loan?'. To the right, it says '\$70,000 Prize Money'. Below the title, there is a red Home Credit Group logo, the text 'Home Credit Group', '7,176 teams', and '3 years ago'. At the bottom, there is a navigation bar with links: 'Overview', 'Data' (which is underlined in blue), 'Code', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and a dark button labeled 'Late Submission' with a three-dot menu icon to its right.


Topics: clients, données personnel


Geo Coverage: Home credit loan

Size: >2Go

 application_test.csv


 application_train.csv

 bureau.csv


 bureau_balance.csv

 credit_card_balance.csv

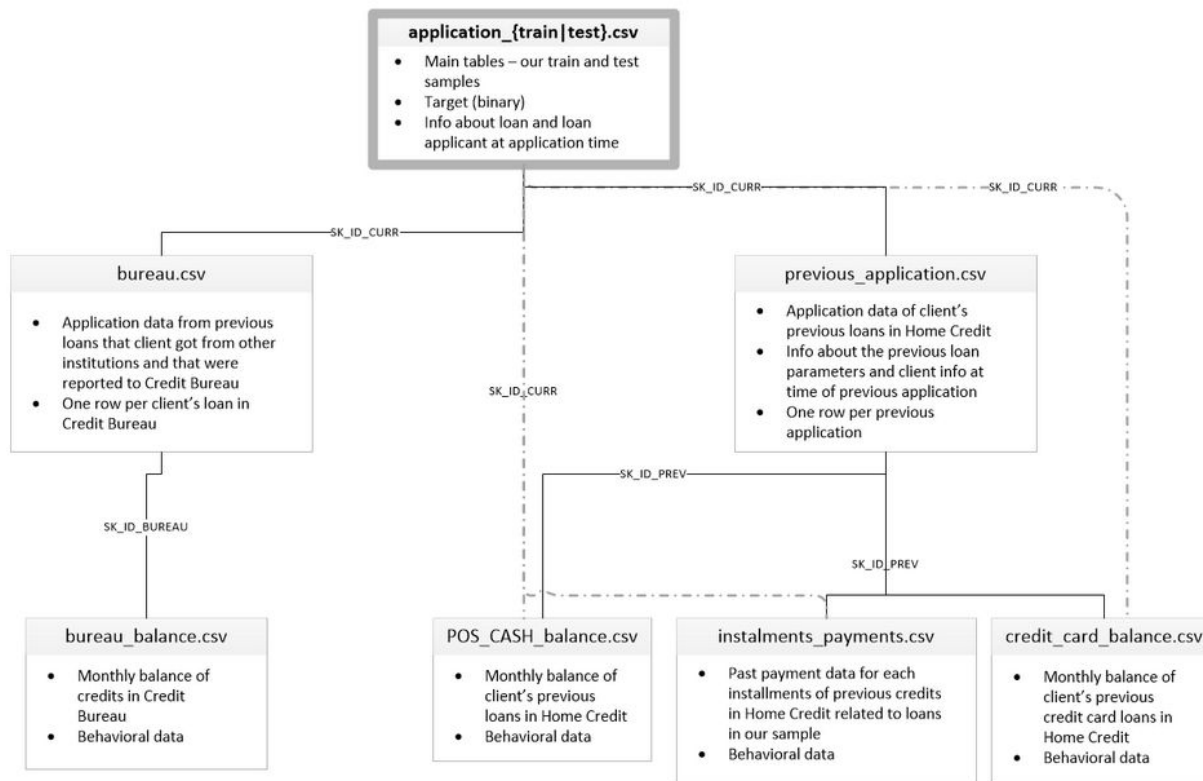
 HomeCredit_columns_description.csv

 installments_payments.csv

 POS_CASH_balance.csv

 previous_application.csv

Les données Les datasets



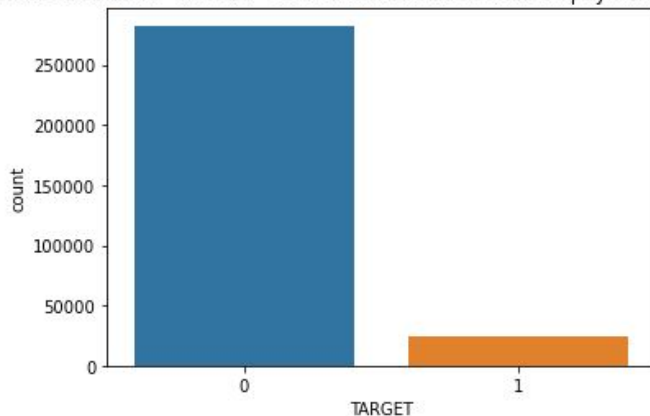
Les données Observation & Compréhension

2 fichiers principaux: application_train.csv et application_test.csv

```
app_train.shape, app_test.shape  
((307511, 122), (48744, 121))
```

Pourcentage de clients en difficultés de payments: 8.072881945686495 %

Distribution de la colonne "TARGET" - 1 -> client avec difficultés de payments / 0 - les autres cas

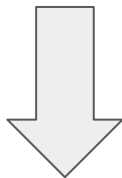


Identification Problème Classification : Non équilibre des classes

92 % peuvent payer, 8% ne peuvent pas



Utilisation de metrics adapté au problèmes: Log loss, Roc auc,F1 score



Choix metriques:

- ROC_AUC (Receiver Operating Characteristic Area Under the Curve)
- F1_score (measure test's accuracy)

Preparation Observations : valeurs manquantes

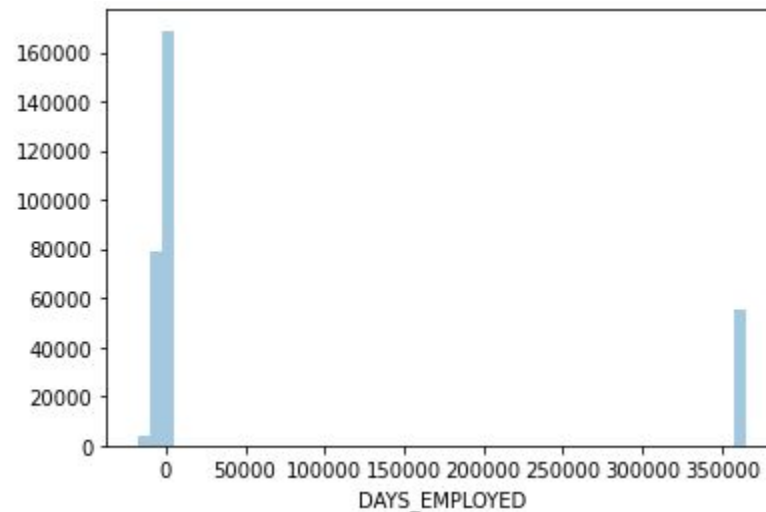
Your selected dataframe has 122 columns.
There are 67 columns that have missing values.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8

Preparation Observations : anomalies

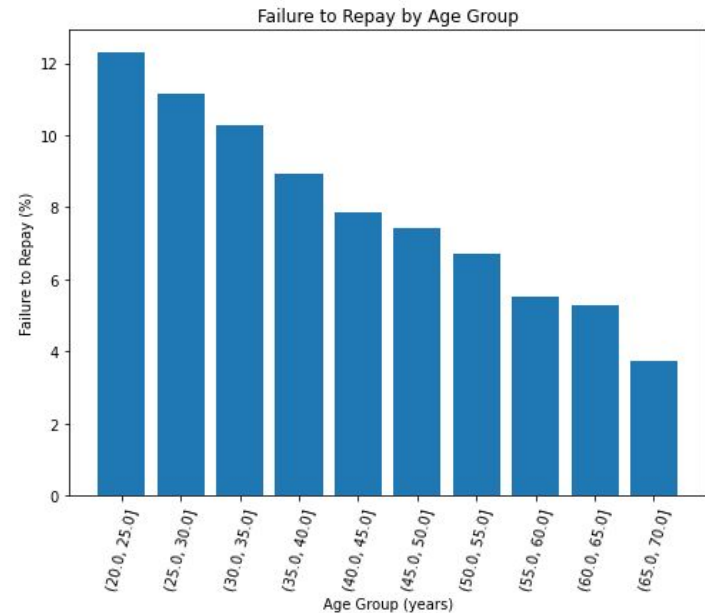
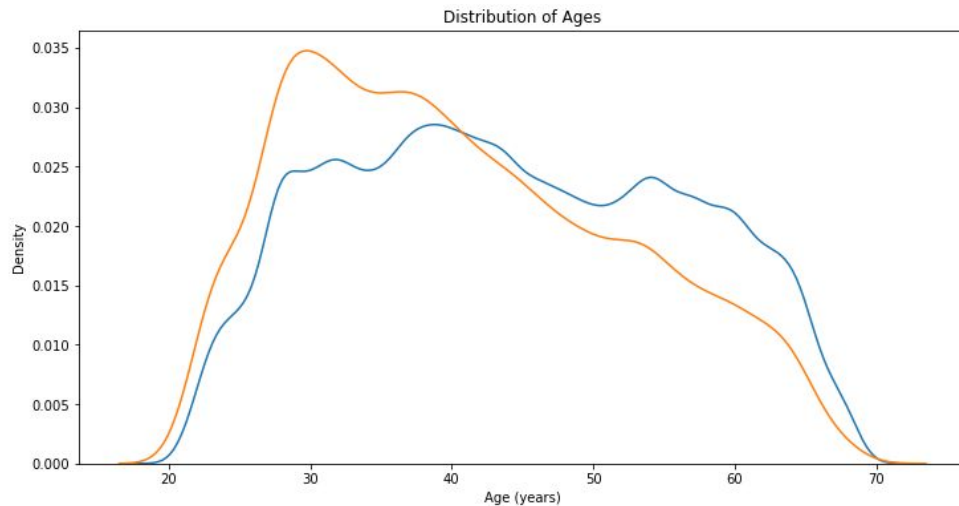
```
app_train['DAYS_EMPLOYED'].describe()
```

```
count    307511.000000
mean      63815.045904
std       141275.766519
min       -17912.000000
25%       -2760.000000
50%       -1213.000000
75%        -289.000000
max       365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```



La valeur maximale est anormal. Elle correspond a 1000 ans, c'est étrange

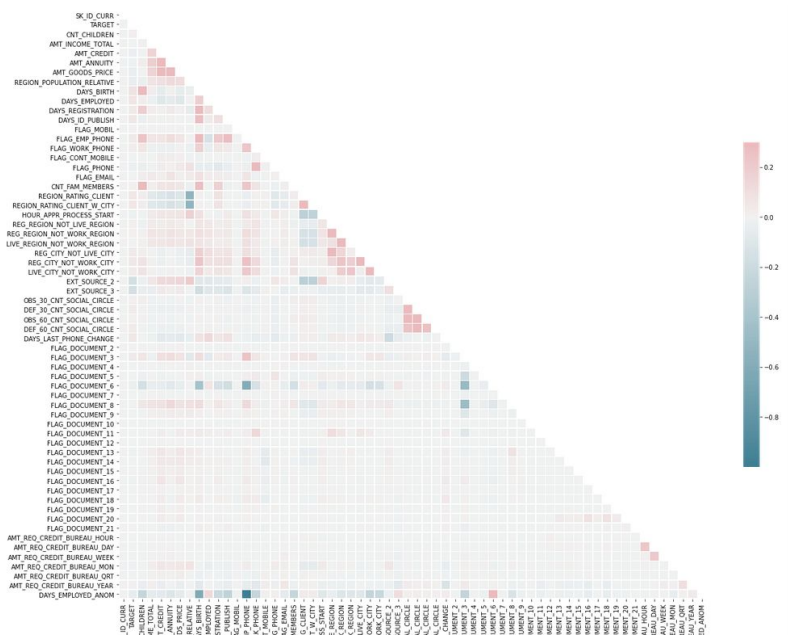
Preparation Observations : Âge et repayment



Les jeunes ne semblent pas aptes à re-payer le prêt

Preparation Observations : corrélations

- Corrélations
-> pas la méthode la plus adapté pour dire l'importance d'une feature, mais peut donner une idée des relations



Preparation Observations : valeurs manquantes

Your selected dataframe has 122 columns.
There are 67 columns that have missing values.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8

Préparation pour utilisation des modèles

- Encodage des données catégoriques
- Alignement des datasets
- Mise à l'échelle des valeurs (scaling)

```
# Number of unique classes in each object column
app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)

: NAME_CONTRACT_TYPE      2
  CODE_GENDER              3
  FLAG_OWN_CAR             2
  FLAG_OWN_REALTY          2
  NAME_TYPE_SUITE          8
  NAME_INCOME_TYPE         8
  NAME_EDUCATION_TYPE      5
  NAME_FAMILY_STATUS       6
  NAME_HOUSING_TYPE        6
  WEEKDAY_APPR_PROCESS_START 7
  ORGANIZATION_TYPE       58
dtype: int64
```

	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EI
0	100002	-0.577538	0.142129	-0.478095	-0.166143	-0.507236	-0.149452	-1.506880	
1	100003	-0.577538	0.426792	1.725450	0.592683	1.600873	-1.252750	0.166821	
2	100004	-0.577538	-0.427196	-1.152888	-1.404669	-1.092145	-0.783451	0.689509	
3	100006	-0.577538	-0.142533	-0.711430	0.177874	-0.653463	-0.928991	0.680114	-
4	100007	-0.577538	-0.199466	-0.213734	-0.361749	-0.068554	0.563570	0.892535	-

Modelisation Algorithmes utilisé

- Type de machine learning
 - Supervisé: labels inclus dans le training data
 - Classification binaire: Target a deux valeurs, 0(peut repayer) et 1(ne pourra pas repayer)

- Baseline: Dummy classifier
- Random forest(ensemble learning)
- LightGBM(decision tree algorithm)
- XGBOOST(decision tree algorithm)

Cross validation, RandomizedSearch



Choix sur metrics

- ROC-AUC
- F1-SCORE

ROC AUC est très utiles lors de prédiction de probabilité sur des résultats binaire, comme notre problème.

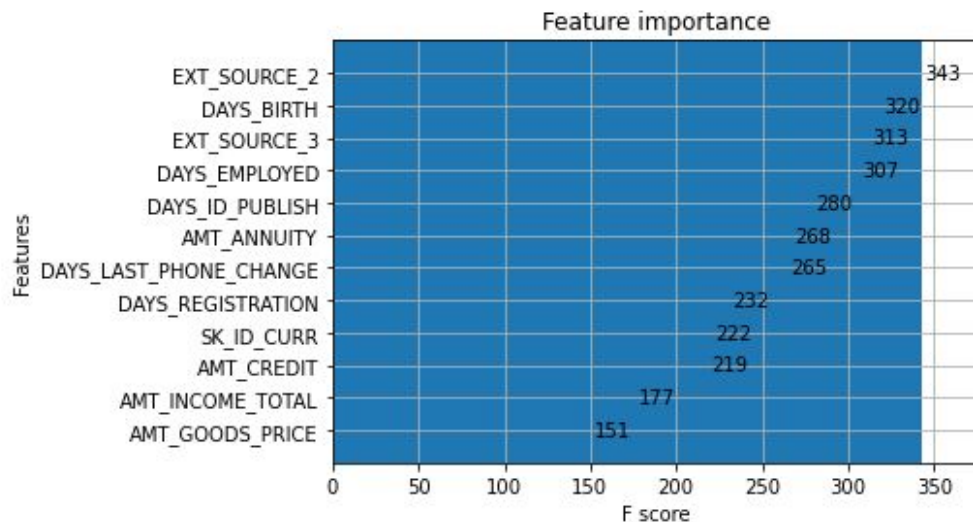
Modelisation Algo utilisé et résultats par metrics

Les résultats sont les suivants:

modèle	Score ROC AUC	F1 Score
Dummy classifier	0.5	0.88
Random Forest	0.5	0.88
LightGBM	0.68	0.76
XGBOOST	0.67	0.80

Les meilleurs modèles sont LightGBM et XGBOOST

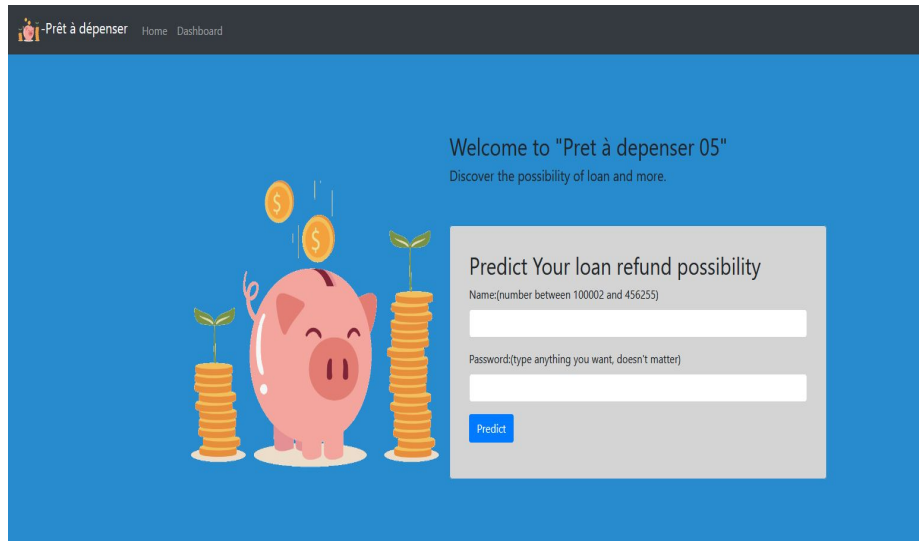
Modelisation Feature importance



API et DASHBOARD

- Api accessible ici: <https://pretadepenser.azurewebsites.net/>
- Dashboard accessible ici: <https://share.streamlit.io/malikhouni/pretadepenser-dash-streamlit/main/app.py>

Host: microsoft azure



Pret à dépenser Home Dashboard

Welcome to "Pret à dépenser 05"
Discover the possibility of loan and more.

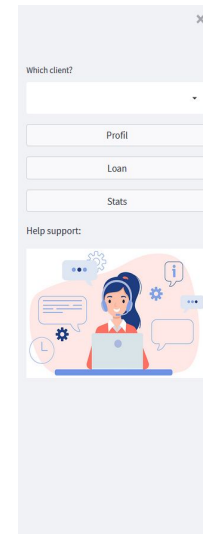
Predict Your loan refund possibility

Name:(number between 100002 and 456255)

Password:(type anything you want, doesn't matter)

Predict

Host: streamlit.io



Dashboard **PRET A DEPENSER**

Please select a Client in the sidebar in the left

Pret à dépenser

Which client?

Profile

Loan

Stats

Help support:

Predict Your loan refund possibility

Name:(number between 100002 and 456255)

Password:(type anything you want, doesn't matter)

Predict

autre

piste de recherches/amélioration possible

- Résumé :
 - Classification binaire
 - Travail sur données importantes(valeurs manquante, anomalies,...)
 - LightGBM et XGboost, bon résultats
 - Hosting à réfléchir

Amélioration pour prochaine études:

- Autre score
- Autre modèle
- Autre hosting solution
- Améliorations de performance du modèle

Fin de présentation
merci de votre attention

- **Sources**

- [Classifiez automatiquement des biens de consommation - OpenClassrooms](https://openclassrooms.com/fr/paths/164/projects/632/assignment)
- <https://openclassrooms.com/fr/paths/164/projects/632/assignment>
- <https://www.kaggle.com/c/home-credit-default-risk/data>