Implémenter un modèle de scoring

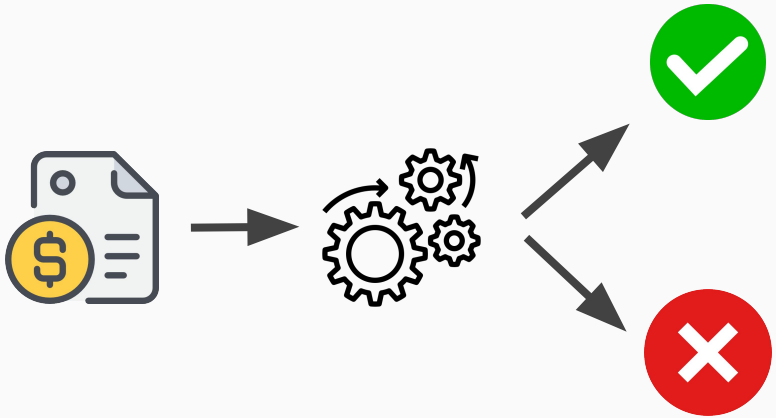
25/05/2022 - Parcours Data Scientist

Sébastien Bourgeois

Sommaire

1. Problématique & dataset

Problématique



Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Problématique

Les icônes de cette slide ont été conçues par Freepik et Smashicons du site Flaticon.com

1. Problématique & dataset

Dataset

Données disponibles

- application_{train|test}.csv
- previous_application.csv

} Données de “Prêt à dépenser”

- bureau.csv
- bureau_balance.csv
- POS_CASH_balance.csv
- credit_card_balance.csv
- installments_payments.csv

1. Problématique & dataset

Dataset

Données disponibles

- application_{train|test}.csv
- previous_application.csv
- bureau.csv
- bureau_balance.csv
- POS_CASH_balance.csv
- credit_card_balance.csv
- installments_payments.csv

Données d'autres institutions financières

1. Problématique & dataset

Dataset

Données disponibles

- application_{train|test}.csv
- previous_application.csv
- bureau.csv
- bureau_balance.csv
- POS_CASH_balance.csv
- credit_card_balance.csv
- installments_payments.csv

} Données comportementales

1. Problématique & dataset

Dataset

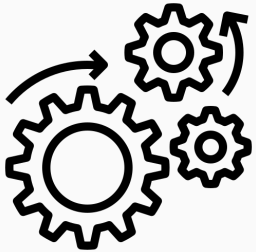
Données disponibles

- application_{train|test}.csv
- previous_application.csv
- bureau.csv
- bureau_balance.csv
- POS_CASH_balance.csv
- credit_card_balance.csv
- installments_payments.csv

1. Problématique & dataset

Dataset

application_train.csv



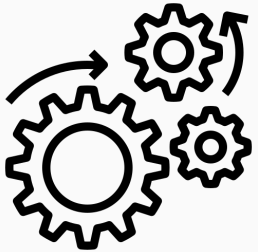
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



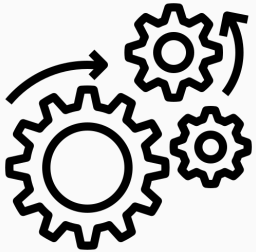
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   SK_ID_CURR                  292813 non-null  int64
 1   TARGET                      292813 non-null  int64
 2   NAME_CONTRACT_TYPE          292813 non-null  object
 3   CODE_GENDER                 292813 non-null  object
 4   FLAG_OWN_CAR                292813 non-null  object
 5   FLAG_OWN_REALTY             292813 non-null  object
 6   AMT_INCOME_TOTAL            292813 non-null  float64
 7   AMT_CREDIT                  292813 non-null  float64
 8   NAME_INCOME_TYPE            292813 non-null  object
 9   NAME_EDUCATION_TYPE         292813 non-null  object
 10  NAME_FAMILY_STATUS          292813 non-null  object
 11  NAME_HOUSING_TYPE           292813 non-null  object
 12  CNT_FAM_MEMBERS             292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE    292813 non-null  float64
 14  CLIENT_AGE                  292813 non-null  float64
 15  OWN_CAR_TYPE                292813 non-null  object
 16  JOB_SENIORITY               292813 non-null  object
 17  ANNUAL_PAYMENT_RATE         292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



Gestion des données manquantes

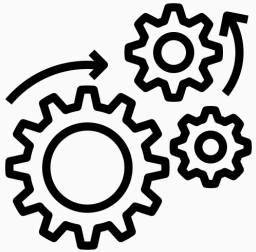
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



Suppression des outliers

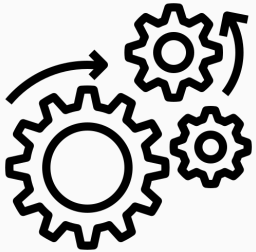
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



Retravail de variables
telle que *CLIENT_AGE*

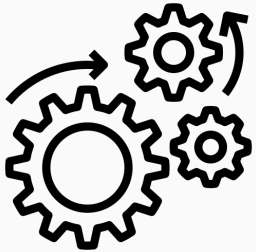
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



Features engineering

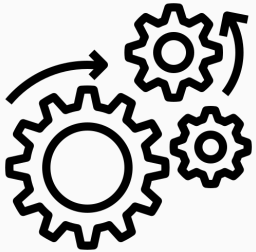
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

1. Problématique & dataset

Dataset

application_train.csv



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292813 entries, 0 to 292812
Data columns (total 18 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   SK_ID_CURR                292813 non-null  int64
 1   TARGET                    292813 non-null  int64
 2   NAME_CONTRACT_TYPE        292813 non-null  object
 3   CODE_GENDER               292813 non-null  object
 4   FLAG_OWN_CAR              292813 non-null  object
 5   FLAG_OWN_REALTY           292813 non-null  object
 6   AMT_INCOME_TOTAL          292813 non-null  float64
 7   AMT_CREDIT                292813 non-null  float64
 8   NAME_INCOME_TYPE          292813 non-null  object
 9   NAME_EDUCATION_TYPE       292813 non-null  object
 10  NAME_FAMILY_STATUS        292813 non-null  object
 11  NAME_HOUSING_TYPE         292813 non-null  object
 12  CNT_FAM_MEMBERS           292813 non-null  float64
 13  DEF_30_CNT_SOCIAL_CIRCLE  292813 non-null  float64
 14  CLIENT_AGE                292813 non-null  float64
 15  OWN_CAR_TYPE              292813 non-null  object
 16  JOB_SENIORITY             292813 non-null  object
 17  ANNUAL_PAYMENT_RATE       292813 non-null  float64
dtypes: float64(6), int64(2), object(10)
memory usage: 40.2+ MB
```

Les icônes de cette slide ont été conçues par justicon, Becris, hqrloveq et Alfredo Hernandez du site Flaticon.com

2. Modélisation

Préparation des données

=>

Dataset (292K lignes)

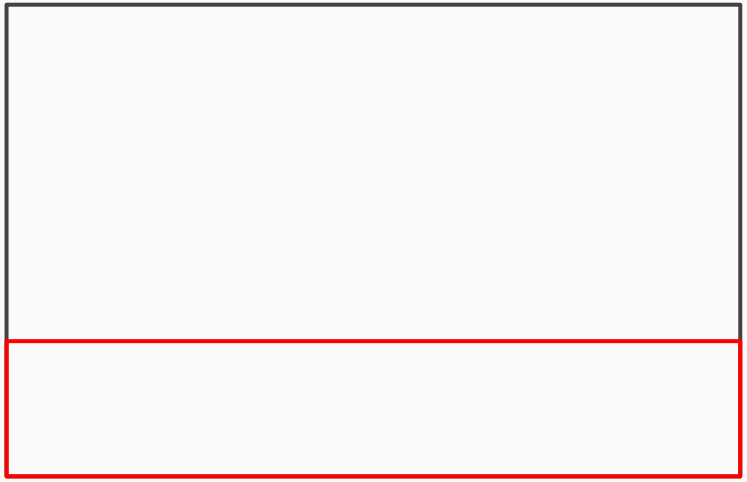
Les icônes de cette slide ont été conçues par nawicon du site Flaticon.com

2. Modélisation

Préparation des données

=>

Dataset (292K lignes)



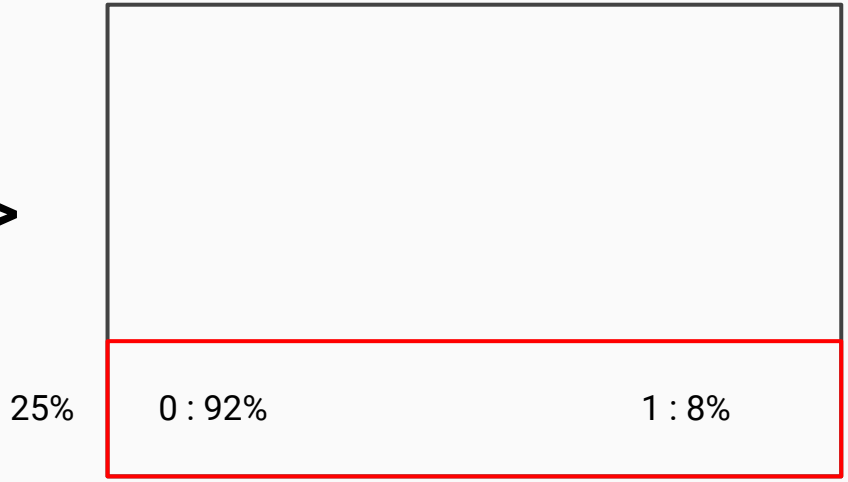
Les icônes de cette slide ont été conçues par nawicon du site Flaticon.com

2. Modélisation

Préparation des données

=>

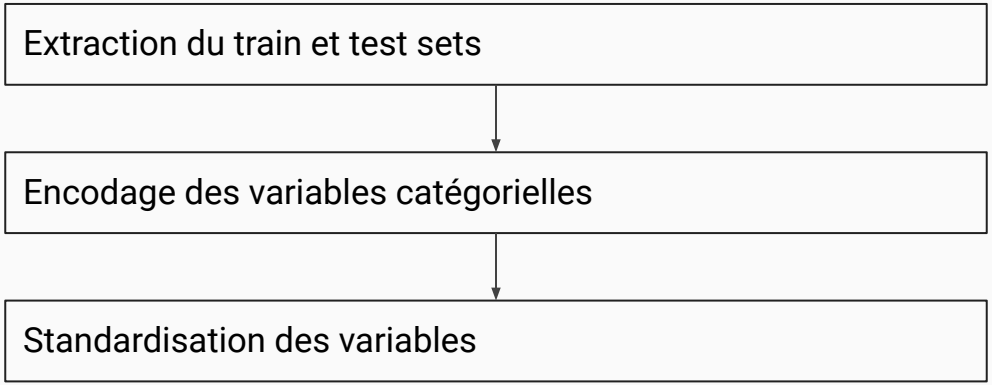
Dataset (292K lignes)



Les icônes de cette slide ont été conçues par nawicon du site Flaticon.com

2. Modélisation

Préparation des données



2. Modélisation

Préparation des données

=>

Dataset (55K lignes)

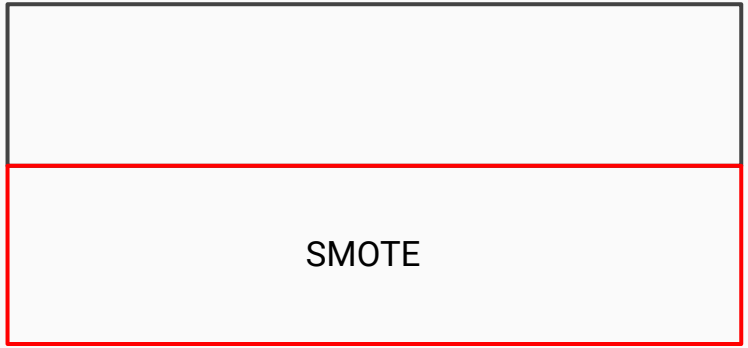
Les icônes de cette slide ont été conçues par Ehtisham Abid du site Flaticon.com

2. Modélisation

Préparation des données

=>

Dataset (55K lignes)



+46K lignes

Les icônes de cette slide ont été conçues par Ehtisham Abid du site Flaticon.com

2. Modélisation

Entraînement des modèles

- Régression logistique

2. Modélisation

Entraînement des modèles

- Régression logistique
- Random Forest

2. Modélisation

Entraînement des modèles

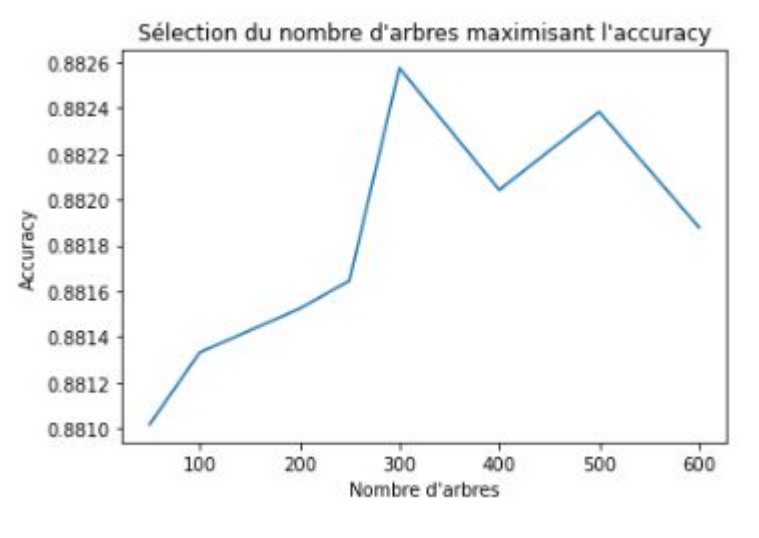
- Régression logistique
- Random Forest
- LightGBM

2. Modélisation

Entraînement des modèles

[50, 100, 200, 250, 300, 400, 500, 600]

- Régression logistique
- Random Forest
- LightGBM

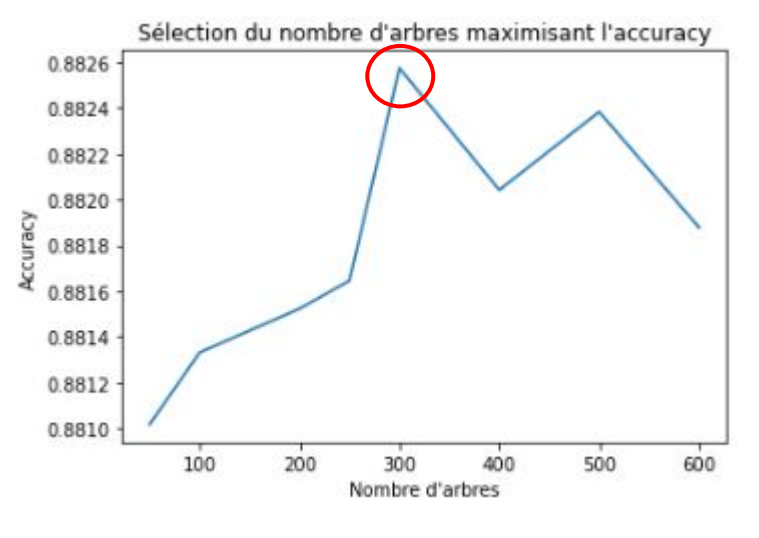


2. Modélisation

Entraînement des modèles

[50, 100, 200, 250, 300, 400, 500, 600]

- Régression logistique
- Random Forest
- LightGBM



2. Modélisation

Entraînement des modèles

- Régression logistique
- Random Forest
- LightGBM

C = [0.001, 0.005, 0.02, 0.1, 0.46, 2.15, 10.0, 46.4, 215.4, 1000.0]

penalty = ['l2', 'none']

2. Modélisation

Entraînement des modèles

- Régression logistique
- Random Forest
- LightGBM

n_estimators = 300

max_features = [10, 15, 20, 30, 40, 48]

max_depth = [10, 15, 20, 25]

2. Modélisation

Entraînement des modèles

- Régression logistique
- Random Forest
- LightGBM

n_estimators = [300, 400, 500, 600, 700]

learning_rate = [0.025, 0.05, 0.1]

max_depth = [1, 2, 3, 4, 5]

2. Modélisation

Entraînement des modèles

- Régression logistique
- Random Forest
- LightGBM

n_estimators = [300, 400, 500, 600, 700]

learning_rate = [0.025, 0.05, 0.1]

max_depth = [1, 2, 3, 4, 5]

2. Modélisation

Entraînement des modèles

Validation croisée

	accuracy	temps
baseline	0.49998	0.018088
regression_logistique	0.623936	0.639016
random_forest	0.937141	131.817862
light_GBM	0.948468	2.563505

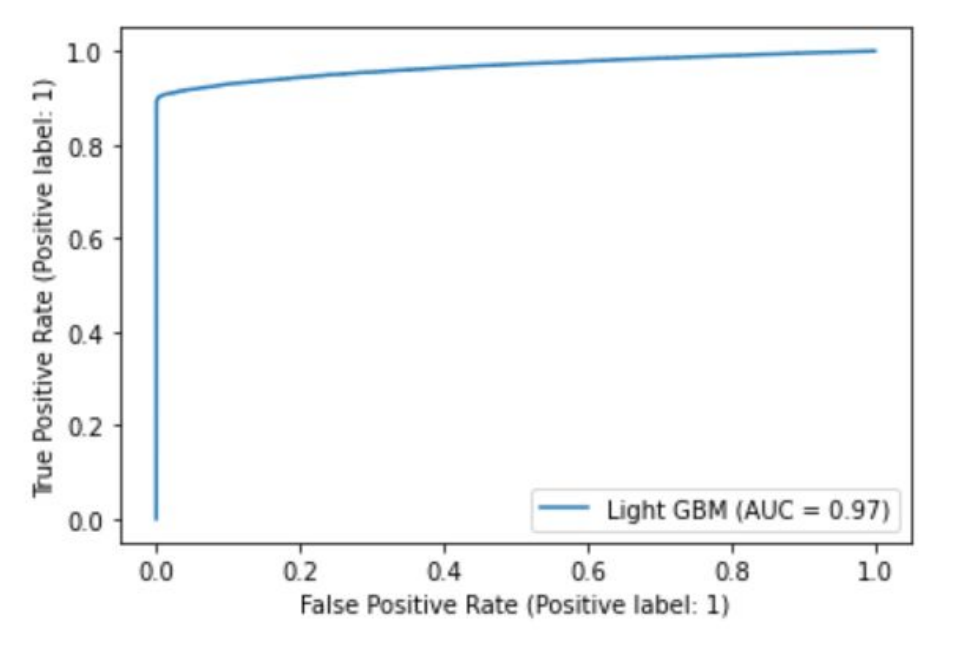
2. Modélisation

Résultats du lightGBM

```
Résultats sur le jeu de test :
- accuracy = 0.9484523809523809
- precision = 0.9915187891440501
- recall = 0.9046428571428572
- fbeta score = 0.920778402481582
```

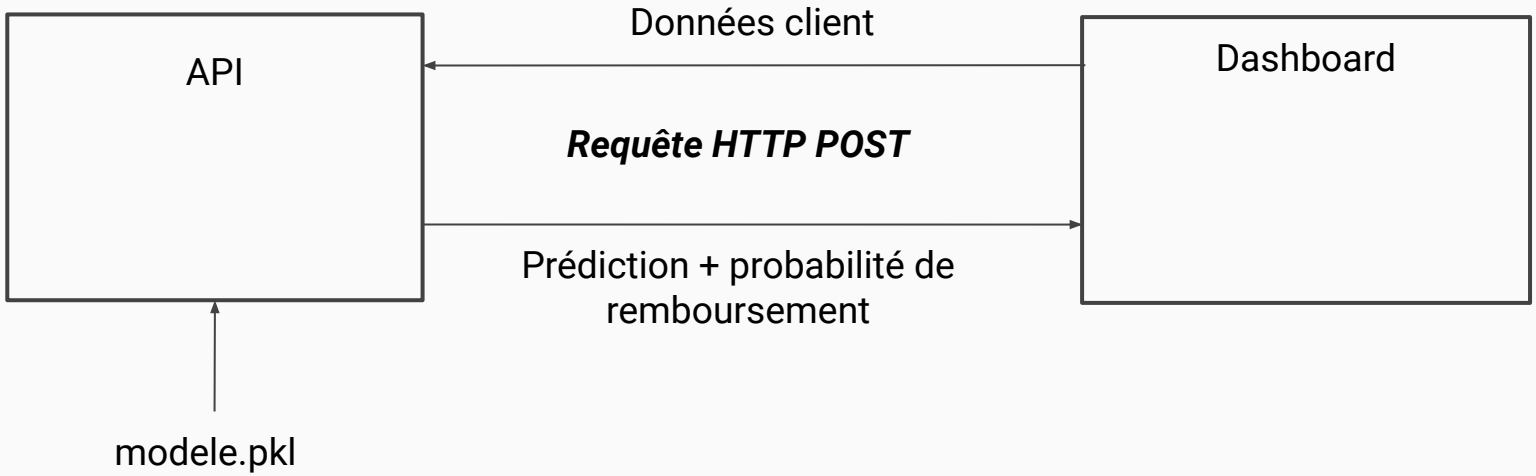
2. Modélisation

Résultats du lightGBM



3. Dashboard

Architecture



3. Dashboard

Liens

Repository Github : https://github.com/sebastienbourgeois/oc-p7-implementer-modele-scoring

API : https://oc-api-modele-scoring.herokuapp.com/

Dashboard : https://oc-dashboard-scoring.herokuapp.com/

4. Conclusion

Déterminer la probabilité de remboursement des clients

4. Conclusion

Déterminer la probabilité de remboursement des clients → LightGBM

4. Conclusion

Déterminer la probabilité de remboursement des clients → LightGBM → Analyser la cause de l'overfitting

Questions/Réponses

Fin.