



Projet 7

Implémentez un modèle de scoring

AL SAMMAN Wassim *Data Scientist Apprenti*

PAPOUTSIS Panayotis *Data Scientist - Mentor*

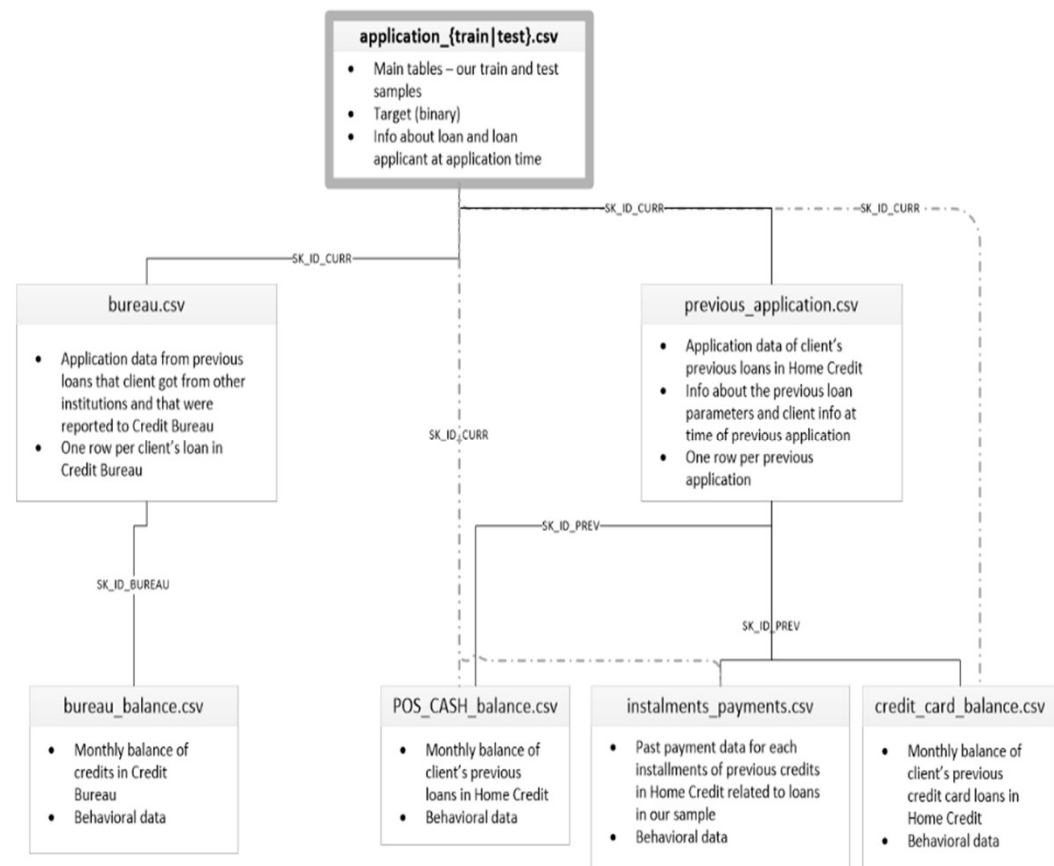
Introduction et problématique

L'entreprise prêt à dépenser souhaite :

- Mettre en œuvre un outil de scoring crédit
- Développer un algorithme de classification
- Développer un Dashboard interactif pour améliorer la transparence de l'entreprise

Présentation du jeu des données

- Taille énorme des données
- Des données test et train
- Combiner les données à l'aide du schéma
- Le TARGET représente les classes



Modélisation_Features engineering

- Les features engineering sur le site Kaggle
- Combiner les résultats
- Traiter les valeurs inf et nan
- Séparer les données entre train et test

Modélisation_Choix des modèles

- Choisir trois modèles de classification :

<code>model_xgb</code>	<code>xgb.XGBClassifier()</code>
<code>model_random_forest</code>	<code>RandomForestClassifier(max_depth=2, random_state=0)</code>
<code>model_gradient_boosting_classifier</code>	<code>GradientBoostingClassifier()</code>

Modélisation_Démarche de modélisation

Entraîner le modèle

Trouver les classes pour X_test

Traiter le déséquilibre

class_weight

SMOTE

Calculer les scores

f1_score

roc_auc_score


Fonction coût métier

Trouver features importances global et local

Modélisation_Fonction coût métier

- Expertise du domaine
- y_{true} et $y_{predicted}$
- La matrice de confusion

	Valeur réelle	Valeur prévue
TP	0	0
TN	1	1
FP	0	1
FN	1	0

 **MÉTRIQUE « MÉTIER »** ⁹

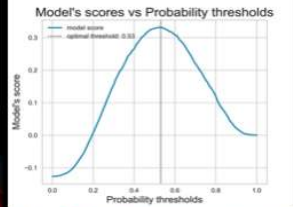
- $gain = TP \cdot TP_value + TN \cdot TN_value + FP \cdot FP_value + FN \cdot FN_value$
- $max_gain = N \cdot TN_value + P \cdot TP_value$
- $baseline = (TN + FP) \cdot TN_value + (TP + FN) \cdot FN_value$

$\Rightarrow score = \frac{gain - baseline}{max_gain - baseline} \in [0; 1]$

$\Rightarrow model_score = \max_{threshold \in [0; 1]} [score] \in [0; 1]$

TP_value = 0
FP_value = 0
FN_value = -10
TN_value = 1

Model's scores vs Probability thresholds



Actual Value
(as confirmed by experiment)

	positives	negatives
positives	TP True Positive	FP False Positive
negatives	FN False Negative	TN True Negative

Predicted Value
(predicted by the test)

Modélisation_Tracking mlflow ui

The screenshot displays the MLflow UI interface for tracking experiments. The 'Experiments' tab is active, and the 'model_random_forest' experiment is selected. The interface shows a list of runs with their respective metrics and parameters.

Experiments

- ☐ Default
- ☐ model_gradient_boosting_cla...
- ☒ model_random_forest
- ☐ model_xgb

model_random_forest

Track machine learning training runs in experiments. Learn more

Experiment ID: 745933057874562526 Artifact Location: http://127.0.0.1:5000

Description Edit

Search: metrics.rmse < 1 and params.model = "tree"

Sort: f1_score Columns

Time created: All time State: Active

Run Name	Created	Duration	Source	Models	Metrics
loud-vole-805	11 hours ago	1.2min	c:\Users\...	-	f1_score: 0.277
aged-bird-203	11 hours ago	33.6s	c:\Users\...	-	f1_score: 0.274
inquisitive-wren-639	11 hours ago	1.9min	c:\Users\...	-	f1_score: 0.228
delightful-wasp-884	11 hours ago	1.1min	c:\Users\...	-	f1_score: 0.066
skillful-dolphin-303	11 hours ago	1.0min	c:\Users\...	-	f1_score: 0.034
enthused-bird-426	11 hours ago	1.1min	c:\Users\...	-	f1_score: 0.033
intrigued-trout-543	11 hours ago	28.7s	c:\Users\...	-	f1_score: 0.015
mercurial-yak-844	11 hours ago	1.9min	c:\Users\...	-	f1_score: 0.013
salty-cow-377	11 hours ago	1.8min	c:\Users\...	-	f1_score: 0.007
traveling-carn-18	11 hours ago	1.3min	c:\Users\...	-	f1_score: 0.006

Show more metrics and parameters (4)

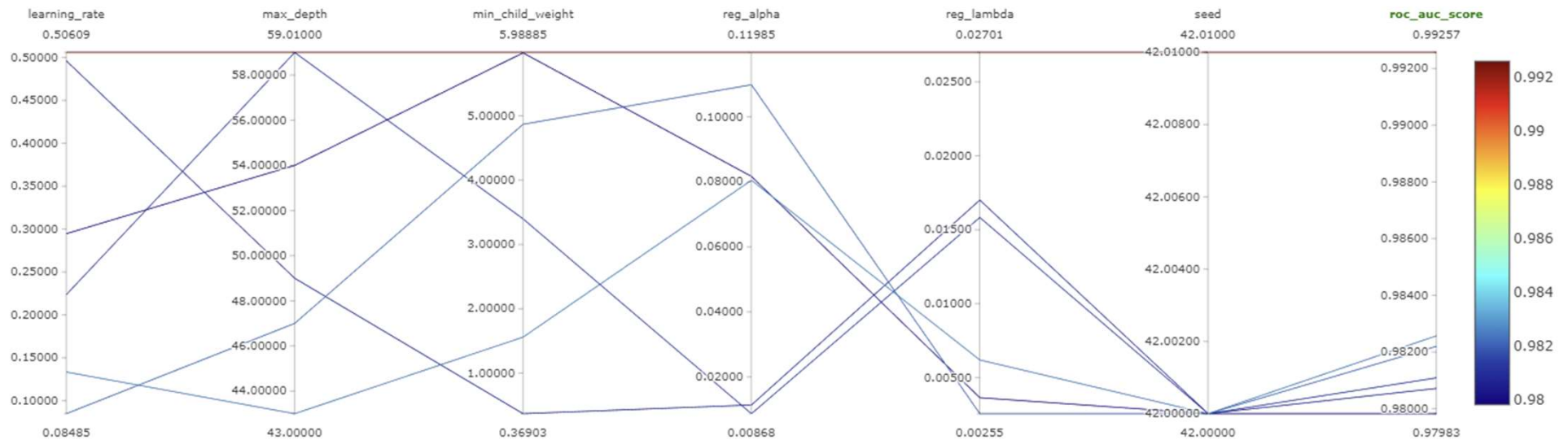
Modélisation_Tracking mlflow ui

Les étapes pour obtenir le tracking présenté dans la photo précédente :

- Définir le tracking uri
- Créer un experiment
- Définir la fonction objective et search space
- Obtenir les paramètres avec la fonction fmin()

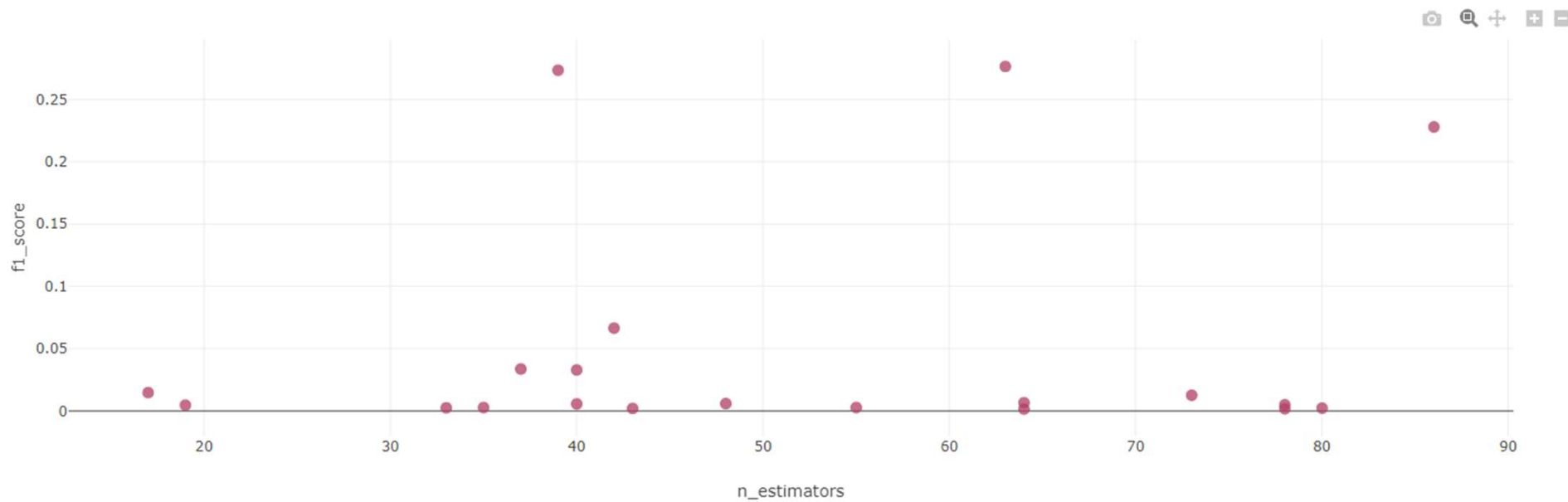
Modélisation_Tracking mlflow ui

- Modèle xgb, les paramètres pour le métrique roc_auc_score



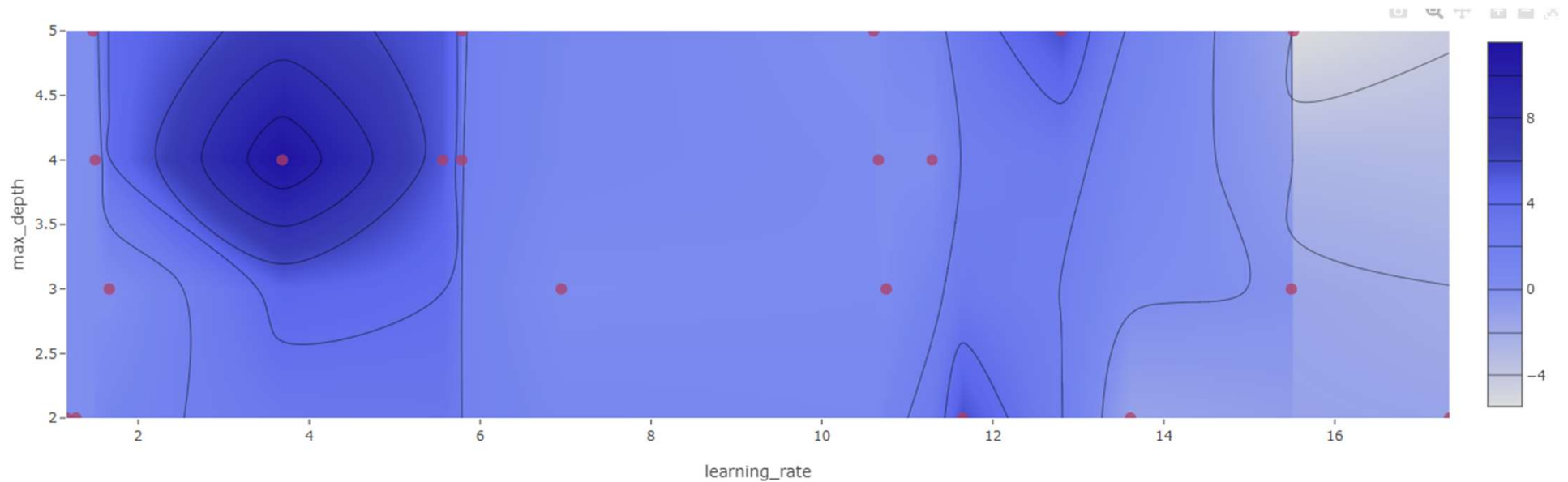
Modélisation_Tracking mlflow ui

- Modèle random_forest, le paramètre n_estimators pour le métrique f1_score



Modélisation_Tracking mlflow ui

- Modèle gradient_boosting_classifier, les paramètres max_depth et learning_rate pour le métrique cost




Modélisation_Synthèses des résultats

		model_xgb	model_random_forest	model_gradient_ boosting_classifier
Sans équilibrer	0	48394	48744	48653
	1	350	0	91
SMOTE	0	48381	39119	48706
	1	363	9625	38
class_weight	0	-	30274	-
	1	-	18470	-
f1_score		0.9535	0.2272	0.0382
roc_auc_score		0.9799	0.7028	0.7695
Coût métier		0.9165	0.4163	0.0180

Modélisation_Synthèses des résultats

Name	Value
learning_rate	0.133401202767
max_depth	43
min_child_weight	1.558998367456
reg_alpha	0.080488258525
reg_lambda	0.006187471945
seed	42

▼ Metrics (1)


Name	Value
roc_auc_score 	0.983

▼ Tags (1)

Name	Value
model	model_xgb

Name	Value
class_weight	balanced
max_depth	13
max_features	sqrt
n_estimators	63

▼ Metrics (1)



Name	Value
f1_score 	0.277

▼ Tags (1)

Name	Value
model	model_random_forest

Name	Value
criterion	squared_error
learning_rate	3.684588876175562
max_depth	4
n_estimators	24

▼ Metrics (2)

Name	Value
cost 	11.53
f1_score 	0.132

▼ Tags (1)

Name	Value
model	model_gradient_boosting_classifier



