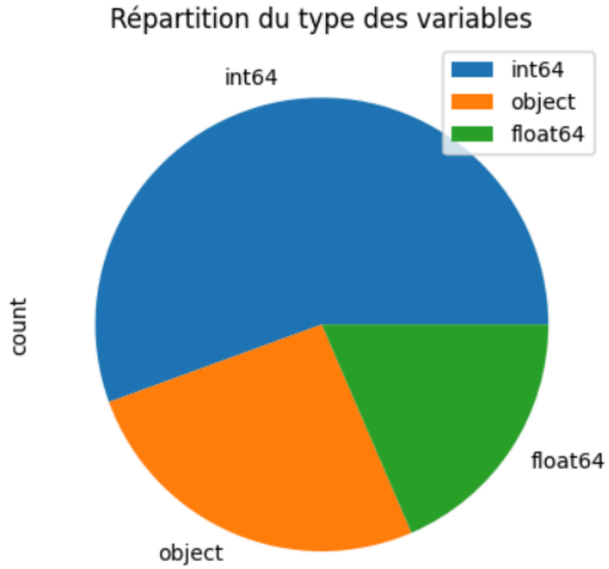

Secure
Trust
Bank

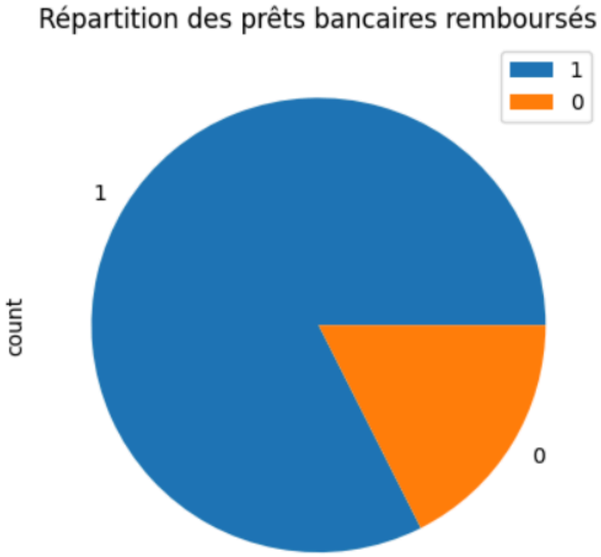
ANALYSE DES DONNEES

Quelques chiffres

Taille du dataset :

- 897 167 individus (lignes) ;
- 27 variables (colonnes) + création de la variable ‘crysis_year’ ;
- La variable cible (Target) : ‘MIS_Status’



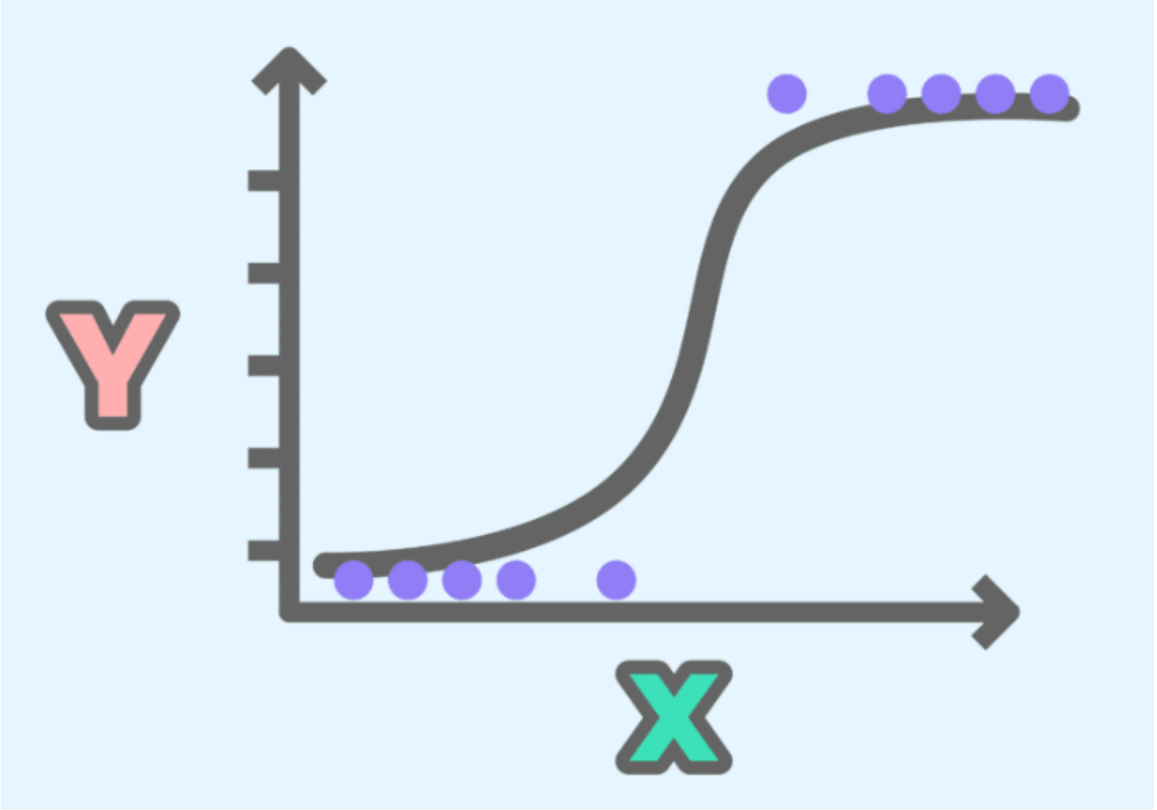


LE MODELE DE REGRESSION LOGISTIQUE

Théorie

<u>Définition</u> :

“Modèle d'analyse statistique *qui permet d'étudier les relations entre un ensemble de variables prédictives nommées X et une variable binomiale nommée Y”.*



Pré-traitement et hyper-paramètres

Pré-traitement des données

Les valeurs manquantes :

- Suppression des variables non-nécessaires, ainsi que celles comprenant plus de 80% de valeurs manquantes ;
- Suppression des lignes avec la valeur de la variable “MIS_Status” nulle ;
- SimpleImputer (imputation par la valeur la + fréquente)

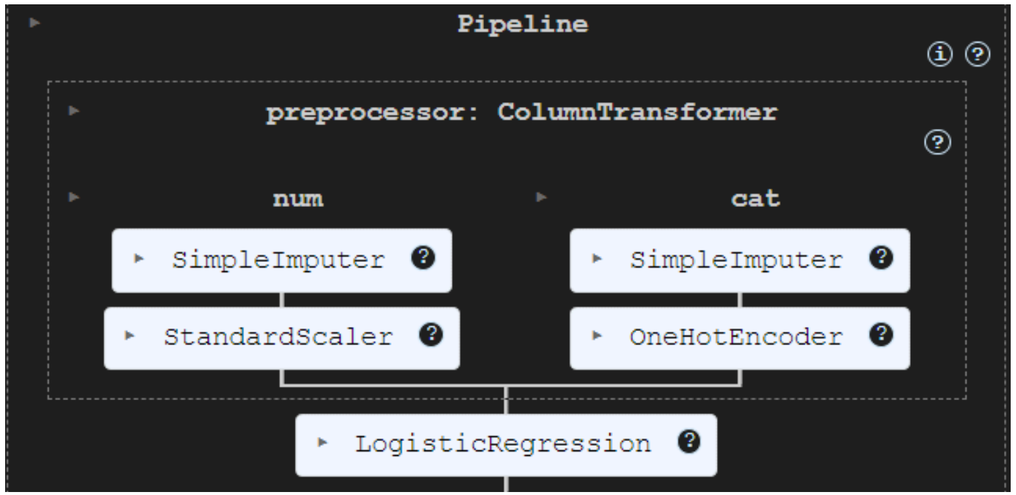
Les valeurs des variables :

- OneHotEncoder (valeurs catégorielles)
- StandardScaler (valeurs numériques)

Hyper-paramètres du modèle

La variable cible (target) :

- pondération des classes prédites (class_weight = ‘balanced’)



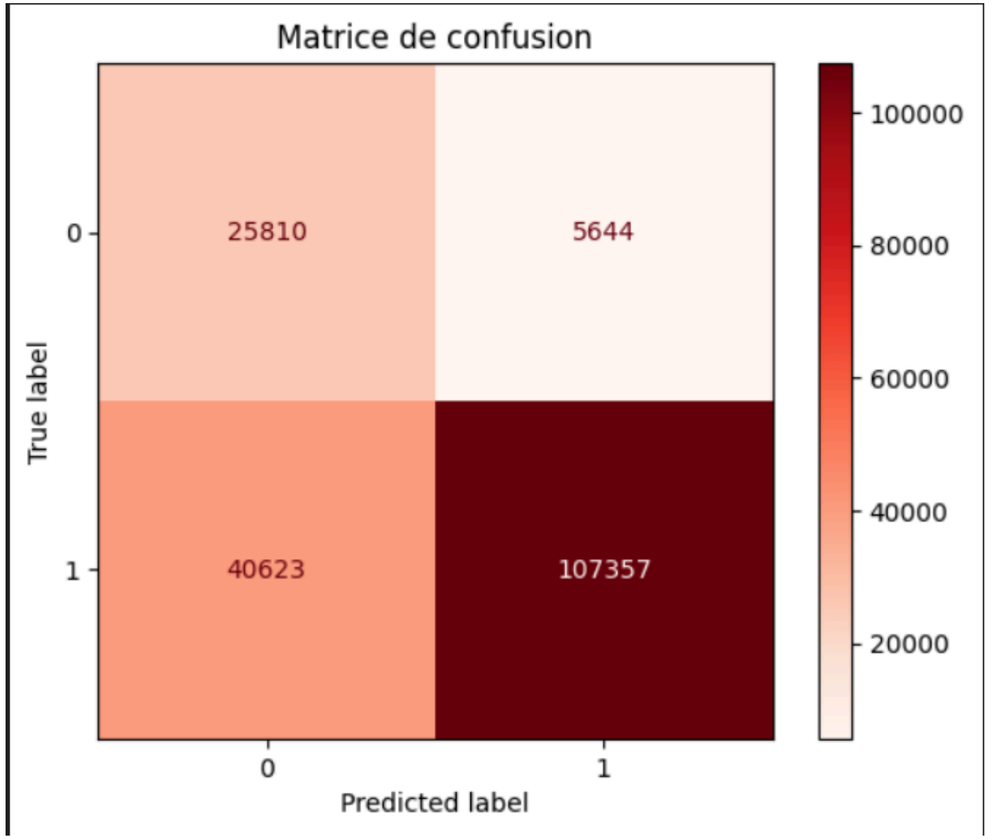
Evaluation du modèle

Score sur les données d'entrainement
Classification Report:

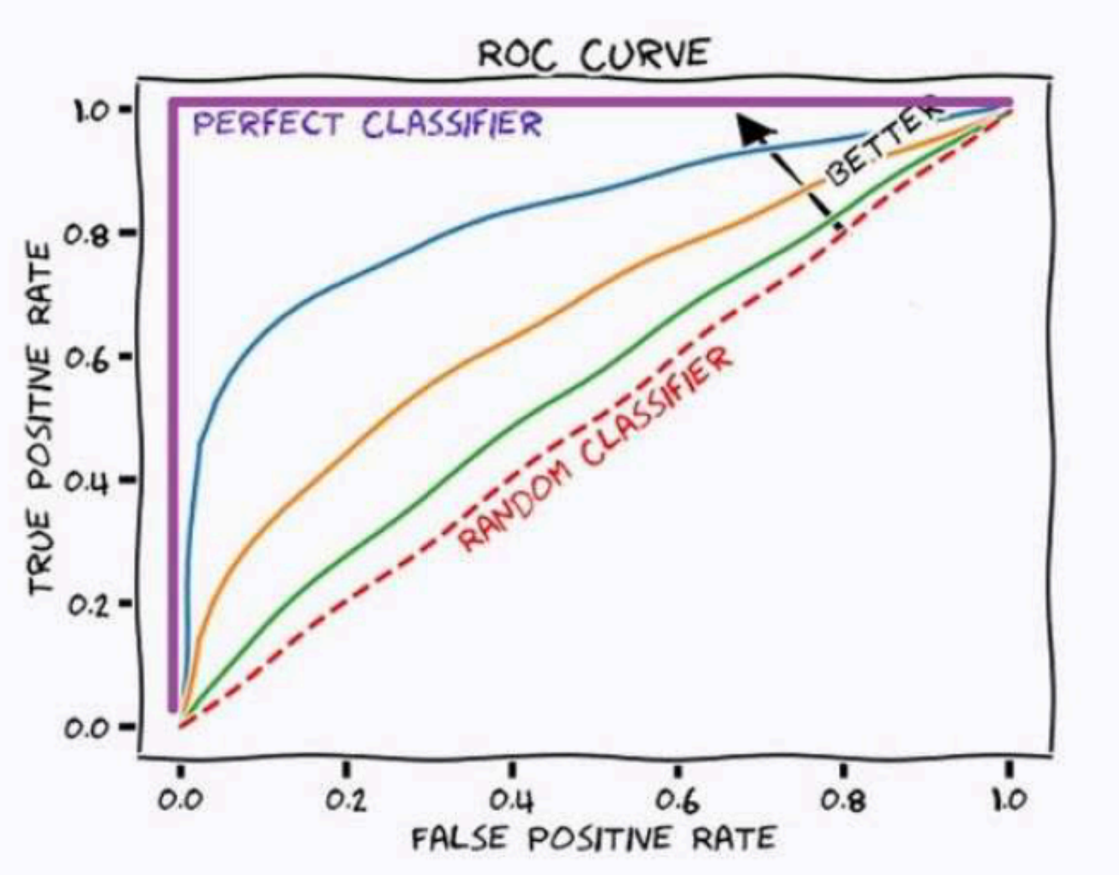
	precision	recall	f1-score	support
0	0.39	0.82	0.53	126104
1	0.95	0.73	0.82	591629
accuracy			0.74	717733
macro avg	0.67	0.77	0.68	717733
weighted avg	0.85	0.74	0.77	717733

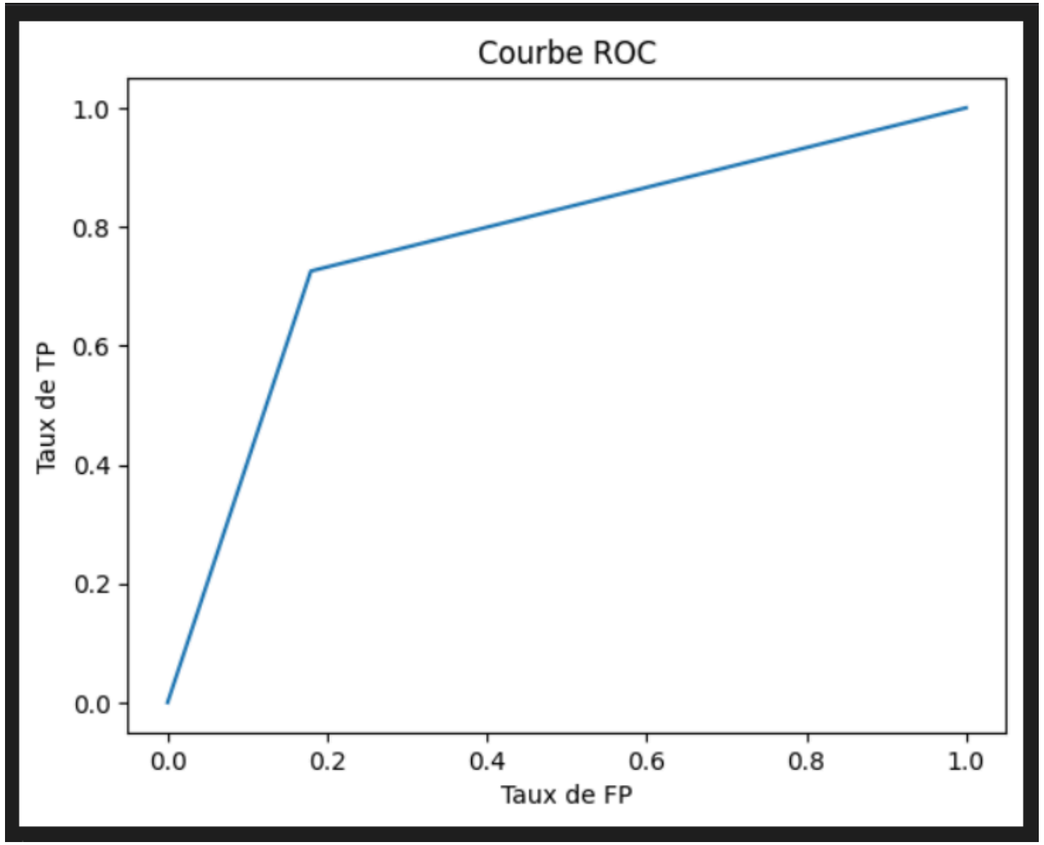
Score sur les données de test
Classification Report:

	precision	recall	f1-score	support
0	0.39	0.82	0.53	31454
1	0.95	0.73	0.82	147980
accuracy			0.74	179434
macro avg	0.67	0.77	0.68	179434
weighted avg	0.85	0.74	0.77	179434



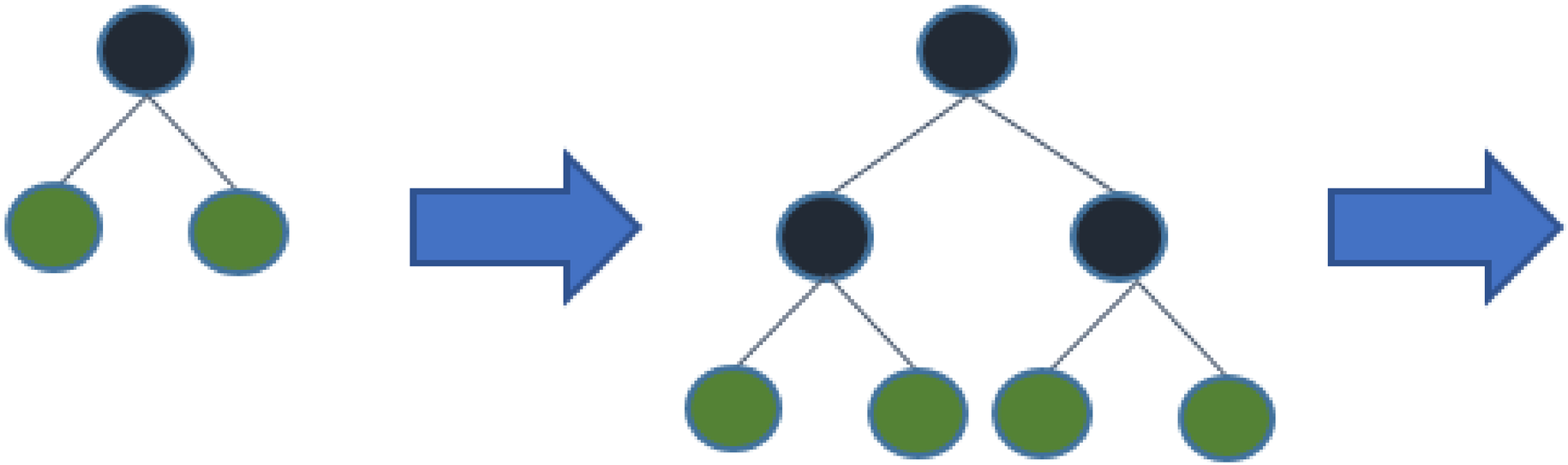
ROC (Reveiver Operating Characteristic)





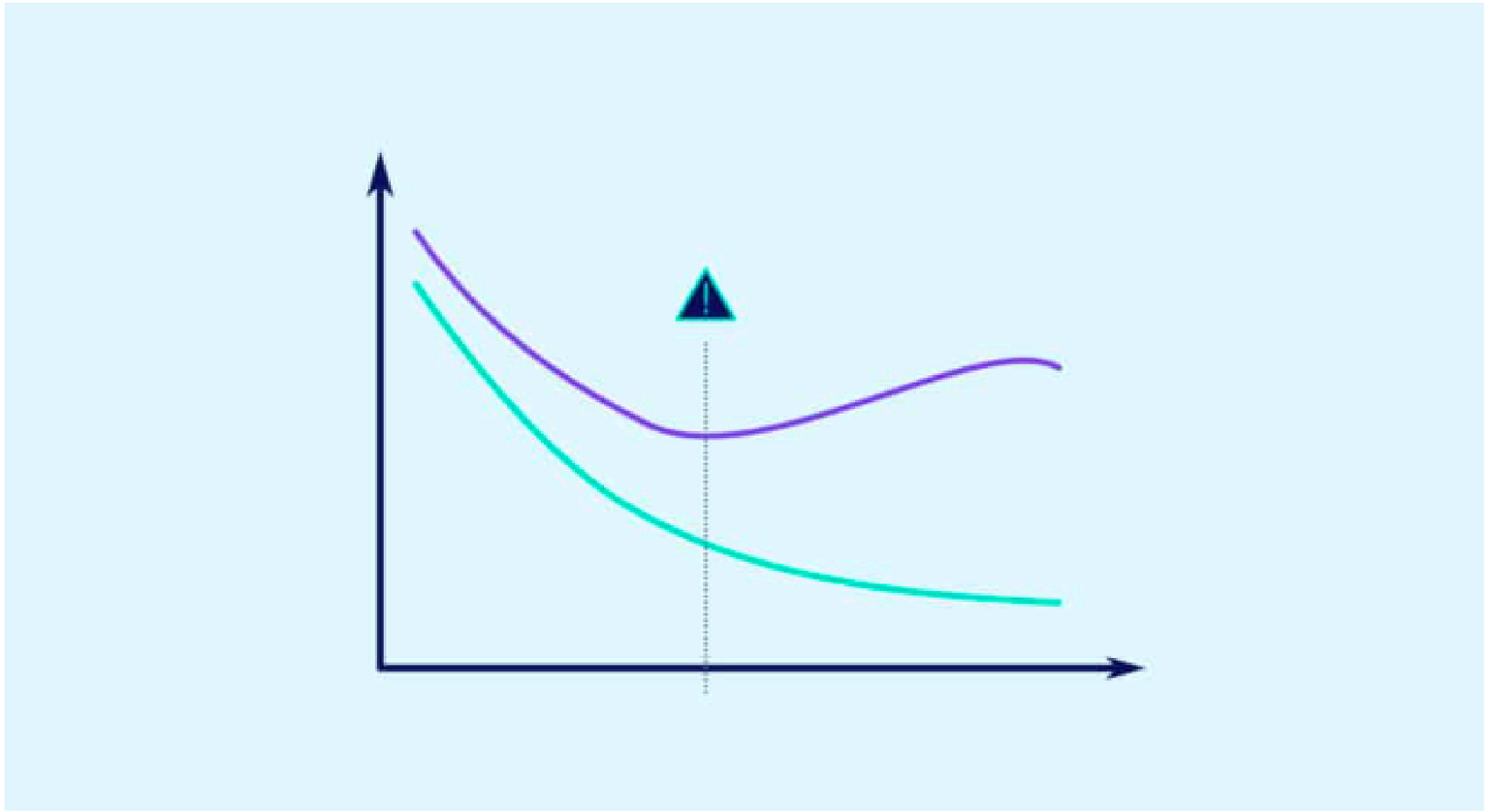
roc_auc = 0.77

LES MODELES DE BOOSTING



Level-wise tree growth

CatBoost pour pallier à l’Overfitting



Pipeline

```
df['State'] = df['State'].where(df['State'].map(df['State'].value_counts()) > 10, 'Other')
```

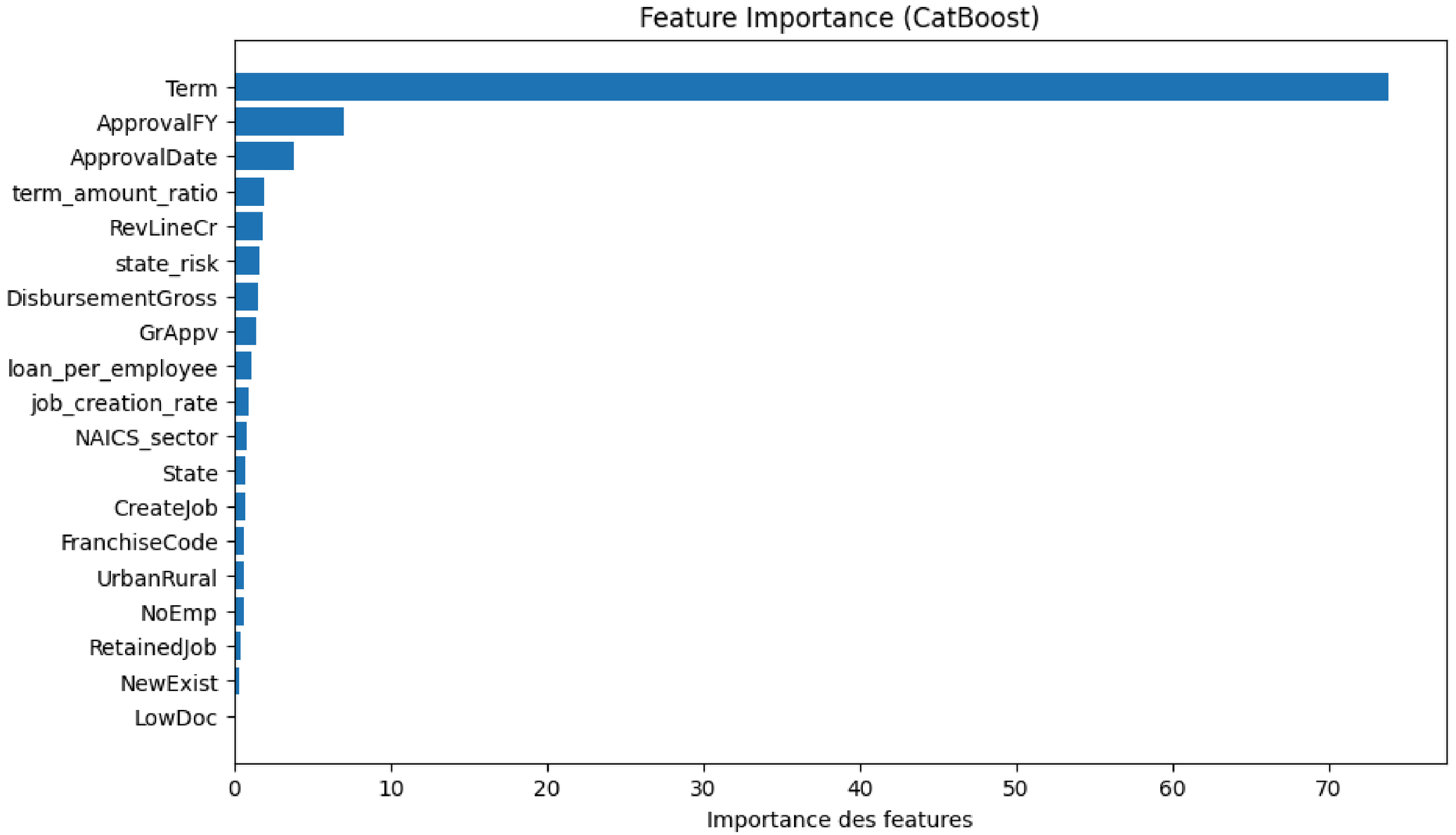
```
def create_advanced_features(df):

    df['term_amount_ratio'] = df['GrAppv'] / df['Term'].replace(0, np.nan)

    state_default_rates = df.groupby('State')['MIS_Status'].mean()
    df['state_risk'] = df['State'].map(state_default_rates)

    df.replace([np.inf, -np.inf], np.nan, inplace=True)

    return df
```

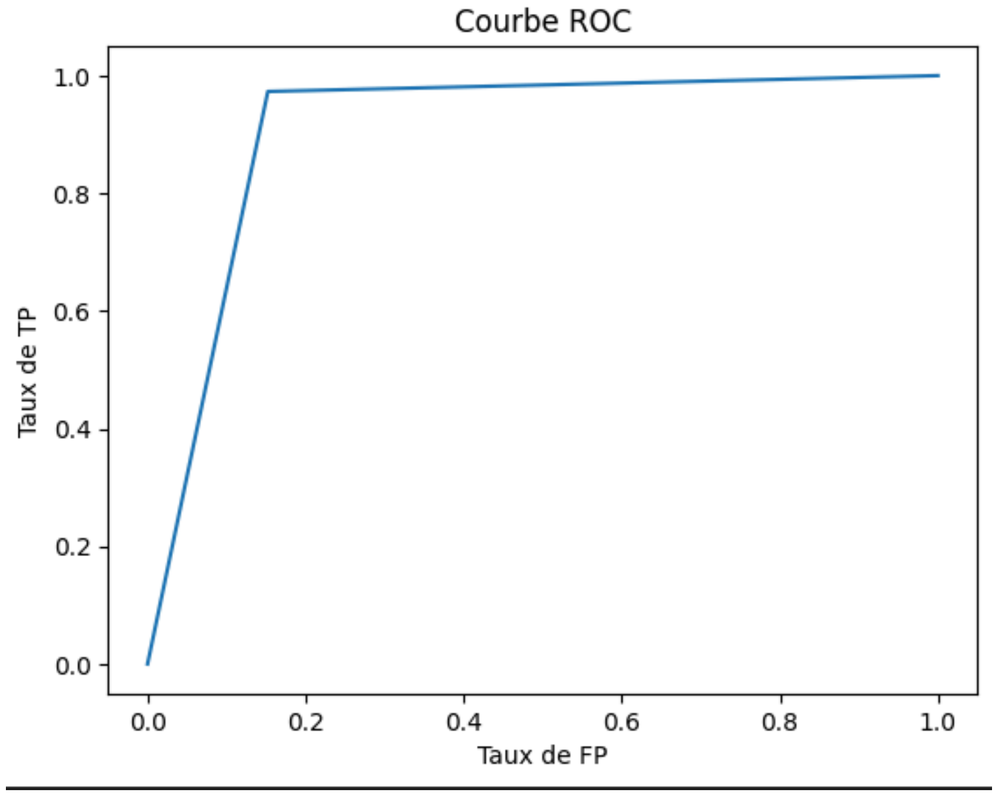
Gridsearch et choix des hyperparamètres choisis

```python
param_grid = {
    'depth': [5],
    'grow_policy': ['Lossguide'],
    'l2_leaf_reg': [0.6],
    'learning_rate': [0.18],
    'scale_pos_weight': [1]
}
```

```python
grid_search = GridSearchCV(
    estimator=CatBoostClassifier(
        iterations=500,
        loss_function='Logloss',
        cat_features=cat_features,
        early_stopping_rounds=50,
        verbose=200
    ),
    param_grid=param_grid,
    cv=3,
    scoring='roc_auc',
    verbose=3,
    n_jobs=-1
)
```

Analyse des résultats





F1 score : 0,95

Recall : 0,95

Precison : 0,95

```
Confusion Matrix:
 [[13292  2464]
 [ 1960 72001]]
```

Comparaison des modèles

LogisticRegression

```
Score sur les données d'entrainement
Classification Report:
              precision    recall  f1-score

           0       0.39      0.82      0.53
           1       0.95      0.73      0.82

    accuracy                           0.74
   macro avg       0.67      0.77      0.68
weighted avg       0.85      0.74      0.77


Score sur les données de test
Classification Report:
              precision    recall  f1-score

           0       0.39      0.82      0.53
           1       0.95      0.73      0.82

    accuracy                           0.74
   macro avg       0.67      0.77      0.68
weighted avg       0.85      0.74      0.77
```

CatBoost

```
📌 Évaluation sur l'ensemble d'entraînement
Classification Report:
              precision    recall  f1-score

           0       0.90      0.88      0.89
           1       0.97      0.98      0.98

    accuracy                           0.96
   macro avg       0.94      0.93      0.93
weighted avg       0.96      0.96      0.96

Confusion Matrix:
 [[110322  15724]
 [ 11966 579721]]

📌 Évaluation sur l'ensemble de test
Classification Report:
              precision    recall  f1-score

           0       0.87      0.85      0.86
           1       0.97      0.97      0.97

    accuracy                           0.95
   macro avg       0.92      0.91      0.91
weighted avg       0.95      0.95      0.95
```