

05_model_training_and_evaluation

September 28, 2024

0.1 Notebook Overview

This notebook builds upon the feature-engineered dataset from the previous notebook (`04_feature_engineering.ipynb`) and focuses on **Model Training and Evaluation**. Our primary goal is to develop a credit risk prediction model that excels at identifying potential loan defaulters, thereby minimizing financial losses for retail banks while also considering their desired balance between risk aversion and loan approval rates. This translates to maximizing the recall of the positive class (loan defaulters) while maintaining acceptable precision and overall model performance.

0.1.1 0.5.1 Objectives

The main objectives of this notebook are:

1. **Model Selection:** Choose algorithms suitable for imbalanced classification problems.
2. **Model Training:** Train models with a focus on identifying potential defaulters.
3. **Hyperparameter Tuning:** Optimize models to increase recall for the positive class.
4. **Model Evaluation:** Assess models primarily on recall, while considering precision, F2-score, AUC-PR, and overall performance.
5. **Model Comparison:** Compare different models based on their ability to identify true positives and balance the precision-recall trade-off.
6. **Threshold Adjustment:** Explore the impact of classification thresholds on recall and precision, collaborating with retail banks to determine the optimal threshold.

0.1.2 0.5.2 Importance of Focusing on Recall

Prioritizing recall for defaulter prediction is crucial for minimizing financial losses, which is the primary business objective in credit risk assessment. The cost of missing a potential defaulter (false negative) is typically much higher than the cost of incorrectly classifying a non-defaulter as high-risk (false positive). While we prioritize recall, we will also carefully consider the precision-recall trade-off and aim for a model that maximizes recall without severely impacting precision. Techniques like threshold adjustment and cost-sensitive learning will be used to balance these metrics effectively. Furthermore, demonstrating a thorough approach to risk identification aligns with regulatory expectations in the financial sector, supporting the banks' compliance needs. This approach also allows for more conservative lending practices, which can be adjusted based on the bank's specific risk tolerance.

0.1.3 0.5.3 Our Approach

In this notebook, we will focus on the following modeling tasks:

1. **Data Preparation:** Address class imbalance using techniques like SMOTE or class weighting.
2. **Baseline Model:** A logistic regression model with class weights inversely proportional to class frequencies will serve as our baseline. This will provide a benchmark for evaluating more complex models.
3. **Advanced Models:** Train and evaluate models known for handling imbalanced data:
 - Decision Trees with adjusted class weights
 - Random Forest with balanced class weights
 - Gradient Boosting (XGBoost, LightGBM) with `scale_pos_weight` adjustment
4. **Hyperparameter Tuning:** We will employ techniques like GridSearchCV or RandomizedSearchCV, optimizing for the F2-score (which gives more weight to recall) or a custom cost-sensitive scoring function.
5. **Model Evaluation:** Prioritize recall in our metrics, while also considering precision, F2-score, AUC-PR, and AUC-ROC.
6. **Threshold Adjustment:** We will experiment with different classification thresholds and work closely with retail banks to determine the optimal threshold that balances their desired level of risk aversion with acceptable loan approval rates.
7. **Ensemble Methods:** Explore ensemble techniques that can improve recall without severely impacting precision.
8. **Cost-Sensitive Learning:** Incorporate misclassification costs to reflect the higher cost of false negatives, aligning the model's objective with the business goal of minimizing financial losses.

By the end of this notebook, we aim to have a model (or ensemble of models) that excels at identifying potential loan defaulters, providing the bank with a powerful tool for risk assessment and mitigation.

```
[8]: import pandas as pd
import numpy as np
import os
import re
from sklearn.preprocessing import FunctionTransformer
from lightgbm.callback import early_stopping

from sklearn.model_selection import (
    train_test_split,
    StratifiedKFold,
    GridSearchCV,
    RandomizedSearchCV,
    cross_val_score,
)
from sklearn.model_selection import cross_val_predict

from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.dummy import DummyClassifier
from sklearn.tree import DecisionTreeClassifier
```

```

from sklearn.pipeline import Pipeline
import xgboost as xgb
import lightgbm as lgb
from sklearn.metrics import (
    classification_report,
    roc_auc_score,
    precision_recall_curve,
    auc,
    confusion_matrix,
    f1_score,
    recall_score,
    precision_score,
    make_scorer,
)
from sklearn.utils.class_weight import compute_class_weight

from retail_bank_risk.model_training_utils import downscale_dtypes
from retail_bank_risk.advanced_visualizations_utils import (
    plot_confusion_matrix,
    plot_model_performance,
    shap_summary_plot,
    shap_force_plot,
    plot_roc_curve,
    plot_precision_recall_curve,
    plot_combined_confusion_matrices,
    plot_learning_curve,
)

from joblib import Parallel, delayed
import warnings
warnings.filterwarnings('ignore') # Suppress warnings for cleaner output

import matplotlib.pyplot as plt
import seaborn as sns
import shap

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, FunctionTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.dummy import DummyClassifier
from sklearn.model_selection import cross_val_score, StratifiedKFold
from sklearn.feature_selection import SelectFromModel

```

```

[2]: train_df = pd.read_parquet("../data/processed/application_train_engineered.
    ↪parquet")

```

```
test_df = pd.read_parquet("../data/processed/application_test_engineered.
↳parquet")
```

```
print(f"Training Data Shape: {train_df.shape}")
```

```
print(f"Test Data Shape: {test_df.shape}")
```

Training Data Shape: (307511, 78)

Test Data Shape: (48744, 77)

```
[3]: train_df, test_df = downscale_dtypes(train_df, test_df, target_column='target')
```

```
train_df.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 307511 entries, 0 to 307510
```

```
Data columns (total 78 columns):
```

#	Column	Non-Null Count	Dtype
0	reg_city_not_work_city_0	307511 non-null	uint8
1	reg_city_not_work_city_1	307511 non-null	uint8
2	region_rating_client_w_city	307511 non-null	float32
3	region_rating_client	307511 non-null	float32
4	name_contract_type_cash loans	307511 non-null	uint8
5	name_contract_type_revolving loans	307511 non-null	uint8
6	code_gender_m	307511 non-null	uint8
7	code_gender_f	307511 non-null	uint8
8	flag_own_car_n	307511 non-null	uint8
9	flag_own_car_y	307511 non-null	uint8
10	flag_own_realty_y	307511 non-null	uint8
11	flag_own_realty_n	307511 non-null	uint8
12	name_type_suite_unaccompanied	307511 non-null	uint8
13	name_type_suite_family	307511 non-null	uint8
14	name_type_suite_spouse, partner	307511 non-null	uint8
15	name_type_suite_children	307511 non-null	uint8
16	name_type_suite_other_a	307511 non-null	uint8
17	name_type_suite_mode	307511 non-null	uint8
18	name_type_suite_other_b	307511 non-null	uint8
19	name_type_suite_group of people	307511 non-null	uint8
20	name_income_type_working	307511 non-null	uint8
21	name_income_type_state servant	307511 non-null	uint8
22	name_income_type_commercial associate	307511 non-null	uint8
23	name_income_type_pensioner	307511 non-null	uint8
24	name_income_type_unemployed	307511 non-null	uint8
25	name_income_type_student	307511 non-null	uint8
26	name_income_type_businessman	307511 non-null	uint8
27	name_income_type_maternity leave	307511 non-null	uint8
28	name_education_type	307511 non-null	float32
29	name_family_status_single / not married	307511 non-null	uint8

30	name_family_status_married	307511	non-null	uint8
31	name_family_status_civil marriage	307511	non-null	uint8
32	name_family_status_widow	307511	non-null	uint8
33	name_family_status_separated	307511	non-null	uint8
34	name_family_status_unknown	307511	non-null	uint8
35	name_housing_type_house / apartment	307511	non-null	uint8
36	name_housing_type_rented apartment	307511	non-null	uint8
37	name_housing_type_with parents	307511	non-null	uint8
38	name_housing_type_municipal apartment	307511	non-null	uint8
39	name_housing_type_office apartment	307511	non-null	uint8
40	name_housing_type_co-op apartment	307511	non-null	uint8
41	occupation_type	307511	non-null	float32
42	weekday_appr_process_start_wednesday	307511	non-null	uint8
43	weekday_appr_process_start_monday	307511	non-null	uint8
44	weekday_appr_process_start_thursday	307511	non-null	uint8
45	weekday_appr_process_start_sunday	307511	non-null	uint8
46	weekday_appr_process_start_saturday	307511	non-null	uint8
47	weekday_appr_process_start_friday	307511	non-null	uint8
48	weekday_appr_process_start_tuesday	307511	non-null	uint8
49	organization_type	307511	non-null	float32
50	housetype_mode_block of flats	307511	non-null	uint8
51	housetype_mode_mode	307511	non-null	uint8
52	housetype_mode_terraced house	307511	non-null	uint8
53	housetype_mode_specific housing	307511	non-null	uint8
54	emergencystate_mode_no	307511	non-null	uint8
55	emergencystate_mode_mode	307511	non-null	uint8
56	emergencystate_mode_yes	307511	non-null	uint8
57	days_last_phone_change	307511	non-null	float32
58	days_birth	307511	non-null	float32
59	days_id_publish	307511	non-null	float32
60	ext_source_3	307511	non-null	float32
61	ext_source_2	307511	non-null	float32
62	sk_id_curr	307511	non-null	float32
63	amt_income_total	307511	non-null	float32
64	amt_credit	307511	non-null	float32
65	amt_annuity	307511	non-null	float32
66	amt_goods_price	307511	non-null	float32
67	is_anomaly_false	307511	non-null	uint8
68	is_anomaly_true	307511	non-null	uint8
69	age_group	307511	non-null	float32
70	income_group	307511	non-null	float32
71	credit_amount_group	307511	non-null	float32
72	debt_to_income_ratio	307511	non-null	float32
73	credit_to_goods_ratio	307511	non-null	float32
74	annuity_to_income_ratio	307511	non-null	float32
75	ext_source_mean	307511	non-null	float32
76	credit_exceeds_goods	307511	non-null	uint8
77	target	307511	non-null	uint8

dtypes: float32(22), uint8(56)
memory usage: 42.2 MB

```
[4]: X = train_df.drop(["target", "sk_id_curr"], axis=1)
y = train_df["target"]

X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2,
↳random_state=42, stratify=y)

X_test = test_df.drop("sk_id_curr", axis=1)
sk_id_curr = test_df["sk_id_curr"]

print(f"Training set shape: {X_train.shape}")
print(f"Validation set shape: {X_val.shape}")
print(f"Test set shape: {X_test.shape}")
```

Training set shape: (246008, 76)
Validation set shape: (61503, 76)
Test set shape: (48744, 76)

```
[5]: def sanitize_feature_names(X):
    return X.rename(columns=lambda x: re.sub(r'[^\\w]+', '_', x))
sanitize_transformer = FunctionTransformer(sanitize_feature_names)
```

```
[6]: pipelines = {
    'Dummy Classifier': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('classifier', DummyClassifier(strategy='stratified', random_state=42))
    ]),
    'Logistic Regression': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('scaler', StandardScaler()),
        ('feature_selection',
↳SelectFromModel(LogisticRegression(random_state=42))),
        ('classifier', LogisticRegression(random_state=42,
↳class_weight='balanced',
max_iter=1000, penalty='l2', C=0.1))
    ]),
    'Decision Tree': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('feature_selection',
↳SelectFromModel(DecisionTreeClassifier(random_state=42))),
        ('classifier', DecisionTreeClassifier(random_state=42,
↳class_weight='balanced',
max_depth=3, min_samples_split=5))
    ]),
    'Random Forest': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
```

```

        ('feature_selection',
SelectFromModel(RandomForestClassifier(random_state=42))),
        ('classifier', RandomForestClassifier(random_state=42,
class_weight='balanced',
n_jobs=1, max_depth=5,
n_estimators=100,
min_samples_split=5,
bootstrap=True))
    ],
    'Gradient Boosting': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('feature_selection',
SelectFromModel(GradientBoostingClassifier(random_state=42))),
        ('classifier', GradientBoostingClassifier(random_state=42, max_depth=3,
n_estimators=100,
learning_rate=0.01,
subsample=0.8,
min_samples_split=5))
    ],
    'XGBoost': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('feature_selection', SelectFromModel(xgb.
XGBClassifier(random_state=42))),
        ('classifier', xgb.XGBClassifier(use_label_encoder=False,
eval_metric='logloss',
random_state=42,
scale_pos_weight=len(y)/sum(y),
max_depth=3, n_estimators=100,
learning_rate=0.01,
subsample=0.8, colsample_bytree=0.8,
min_child_weight=5, n_jobs=1))
    ],
    'LightGBM': Pipeline([
        ('sanitizer', FunctionTransformer(sanitize_feature_names)),
        ('feature_selection', SelectFromModel(lgb.
LGBMClassifier(random_state=42))),
        ('classifier', lgb.LGBMClassifier(random_state=42,
class_weight='balanced',
max_depth=3, n_estimators=100,
learning_rate=0.01,
subsample=0.8, colsample_bytree=0.8,
min_child_samples=5, n_jobs=1))
    ])
}

```

```
[9]: def evaluate_model(name, model, X, y):
    print(f"Evaluating {name}...")
    cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

    # Predictions
    y_pred = cross_val_predict(model, X, y, cv=cv, method='predict')
    y_pred_proba = cross_val_predict(model, X, y, cv=cv,
    method='predict_proba')[:, 1]

    # Metrics
    precision = precision_score(y, y_pred)
    recall = recall_score(y, y_pred)
    f1 = f1_score(y, y_pred)
    auc_roc = roc_auc_score(y, y_pred_proba)

    print(f"{name} Cross-validation results:")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1-Score: {f1:.4f}")
    print(f"AUC-ROC: {auc_roc:.4f}")
    print("=" * 60 + "\n")

    return {
        'model': name,
        'precision': precision,
        'recall': recall,
        'f1_score': f1,
        'auc_roc': auc_roc
    }

# Evaluate models
results = []
for name, pipeline in pipelines.items():
    result = evaluate_model(name, pipeline, X, y)
    results.append(result)

# Sort and print final ranking
print("Model Performance Ranking:")
for metric in ['precision', 'recall', 'f1_score', 'auc_roc']:
    print(f"\nRanking by {metric}:")
    sorted_results = sorted(results, key=lambda x: x[metric], reverse=True)
    for i, result in enumerate(sorted_results, 1):
        print(f"{i}. {result['model']}: {metric} = {result[metric]:.4f}")
```

```
Evaluating Dummy Classifier...
Dummy Classifier Cross-validation results:
Precision: 0.0831
Recall: 0.0822
```


F1-Score: 0.0826
AUC-ROC: 0.5013

=====

Evaluating Logistic Regression...
Logistic Regression Cross-validation results:
Precision: 0.1566
Recall: 0.6663
F1-Score: 0.2536
AUC-ROC: 0.7372

=====

Evaluating Decision Tree...
Decision Tree Cross-validation results:
Precision: 0.2039
Recall: 0.6926
F1-Score: 0.3151
AUC-ROC: 0.7981

=====

Evaluating Random Forest...
Random Forest Cross-validation results:
Precision: 0.2550
Recall: 0.7633
F1-Score: 0.3822
AUC-ROC: 0.8873

=====

Evaluating Gradient Boosting...
Gradient Boosting Cross-validation results:
Precision: 1.0000
Recall: 0.1291
F1-Score: 0.2287
AUC-ROC: 0.8503

=====

Evaluating XGBoost...
XGBoost Cross-validation results:
Precision: 0.3213
Recall: 0.8750
F1-Score: 0.4700
AUC-ROC: 0.9448

=====

Evaluating LightGBM...
[LightGBM] [Info] Number of positive: 19860, number of negative: 226148
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.017542 seconds.

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 3535
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432482
```

```
[LightGBM] [Info] Number of positive: 19860, number of negative: 226148
```

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 407
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

[illegible]

```

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.013664 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3542
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 72
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
[LightGBM] [Info] Start training from score -2.432486
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.000837 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 154
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 2
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
[LightGBM] [Info] Start training from score -0.000000
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

```

[illegible]

[illegible]

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 3530
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
```

```
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
```

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 407
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain. best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain. best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
```

[illegible]


```

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.018620 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3529
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 71
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
[LightGBM] [Info] Start training from score -2.432486
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.000935 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 407
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 3
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
[LightGBM] [Info] Start training from score -0.000000
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

```

[illegible]

[illegible]

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 3537
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
```

```
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
```

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 407
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with negative gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
```

[illegible]

```

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Info] Number of positive: 19860, number of negative: 226148
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.018129 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3535
[LightGBM] [Info] Number of data points in the train set: 246008, number of used
features: 72
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432482
[LightGBM] [Info] Start training from score -2.432482
[LightGBM] [Info] Number of positive: 19860, number of negative: 226148
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.000948 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 407
[LightGBM] [Info] Number of data points in the train set: 246008, number of used
features: 3
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
[LightGBM] [Info] Start training from score -0.000000
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

```

[illegible]

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 3542
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
```

```
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
```

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 154
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[EightGBM] [warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

[illegible]

```

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.016368 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3530
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 71
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
[LightGBM] [Info] Start training from score -2.432486
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.000979 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 407
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 3
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
[LightGBM] [Info] Start training from score -0.000000
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

```

[illegible]

[illegible]

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 3529
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
```

```
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
```

You can set ``force_row_wise=true`` to remove the overhead.

```
[LightGBM] [Info] Total Bins 407
```

```
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with negative gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
```

[illegible]

```

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.014930 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3537
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 72
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.080729 -> initscore=-2.432486
[LightGBM] [Info] Start training from score -2.432486
[LightGBM] [Info] Number of positive: 19860, number of negative: 226149
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of
testing was 0.000939 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 407
[LightGBM] [Info] Number of data points in the train set: 246009, number of used
features: 3
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=-0.000000
[LightGBM] [Info] Start training from score -0.000000
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

```


[illegible]

F1-Score: 0.4383

AUC-ROC: 0.9262

=====

Model Performance Ranking:

Ranking by precision:

1. Gradient Boosting: precision = 1.0000
2. XGBoost: precision = 0.3213
3. LightGBM: precision = 0.2963
4. Random Forest: precision = 0.2550
5. Decision Tree: precision = 0.2039
6. Logistic Regression: precision = 0.1566
7. Dummy Classifier: precision = 0.0831

Ranking by recall:

1. XGBoost: recall = 0.8750
2. LightGBM: recall = 0.8415
3. Random Forest: recall = 0.7633
4. Decision Tree: recall = 0.6926
5. Logistic Regression: recall = 0.6663
6. Gradient Boosting: recall = 0.1291
7. Dummy Classifier: recall = 0.0822

Ranking by f1_score:

1. XGBoost: f1_score = 0.4700
2. LightGBM: f1_score = 0.4383
3. Random Forest: f1_score = 0.3822
4. Decision Tree: f1_score = 0.3151
5. Logistic Regression: f1_score = 0.2536
6. Gradient Boosting: f1_score = 0.2287
7. Dummy Classifier: f1_score = 0.0826

Ranking by auc_roc:

1. XGBoost: auc_roc = 0.9448
2. LightGBM: auc_roc = 0.9262
3. Random Forest: auc_roc = 0.8873
4. Gradient Boosting: auc_roc = 0.8503
5. Decision Tree: auc_roc = 0.7981
6. Logistic Regression: auc_roc = 0.7372
7. Dummy Classifier: auc_roc = 0.5013