Tuning Models

На данном этапе загружается ранее сохранённая baseline-модель логистической регрессии вместе с объектом предобработки и метриками качества. Также подгружаются обучающая и тестовая выборки, полученные на этапе предобработки данных.

Это позволяет продолжить работу с моделью, не переобучая её заново, и использовать единый пайплайн обработки признаков.

In [49]:
```
import joblib
import optuna
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score, f1_score, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import os
from catboost import CatBoostClassifier
import pandas as pd
```

Загружаем нашу базовую модель(Logistic regression)

In [ ]:
```
artifact = joblib.load("models/lr_baseline_model.joblib")
lr_baseline = artifact["model"]
preprocessor = artifact["preprocessor"]
baseline_metrics = artifact["metrics"]
```

In [3]:
```
data = joblib.load("lr_preprocessing.joblib")
```

Загрузка пердобработанных данных.

In [ ]:
```
X_train = data["X_train"]
X_test  = data["X_test"]
y_train = data["y_train"]
y_test  = data["y_test"]
```

In [ ]:
```
X_train_lr = preprocessor.fit_transform(X_train)
X_test_lr  = preprocessor.transform(X_test)
X_train_lr.shape, X_test_lr.shape
```

Out[ ]:
```
((800, 52), (200, 52))
```

In [7]:
```
print("NaN in train:", np.isnan(X_train_lr).sum())
print("NaN in test :", np.isnan(X_test_lr).sum())
```

```
NaN in train: 0
NaN in test : 0
```

Подготовка функции оценки качества модели

Для унификации оценки качества классификации создаётся отдельная функция, которая рассчитывает ключевые метрики и визуализирует матрицу ошибок.

Функция позволяет:

- учитывать вероятностные предсказания
- явно задавать порог классификации
- использовать единый подход для всех моделей

In [11]:
```python
def evaluate_classification_model(
    y_true,
    y_proba,
    threshold=0.5,
    model_name="Model"
):
    """
    Универсальная функция оценки бинарной классификации

    Parameters:
    y_true : array-like
        Истинные метки (0/1)
    y_proba : array-like
        Вероятности класса 1
    threshold : float, default=0.5
        Порог классификации
    model_name : str
        Название модели (для визуализаций)

    Returns:
    dict with ROC-AUC and F1-score
    """
    y_pred = (y_proba >= threshold).astype(int)
    roc_auc = roc_auc_score(y_true, y_proba)
    f1 = f1_score(y_true, y_pred)
    print(f"{model_name}")
    print("-" * len(model_name))
    print(f"ROC-AUC : {roc_auc:.4f}")
    print(f"F1-score: {f1:.4f}")
    print(f"Threshold: {threshold}")
    cm = confusion_matrix(y_true, y_pred)
    plt.figure(figsize=(5, 4))
    sns.heatmap(
        cm,
        annot=True,
        fmt="d",
        cmap="Blues",
        cbar=False
    )
    plt.xlabel("Predicted label")
    plt.ylabel("True label")
    plt.title(f"Confusion Matrix — {model_name}")
    plt.show()
    return {
        "model": model_name,
        "roc_auc": roc_auc,
```

```
        "f1_score": f1,
        "threshold": threshold
    }
```

In [19]:
```
X_tr, X_val, y_tr, y_val = train_test_split(
    X_train_lr,
    y_train,
    test_size=0.25,
    stratify=y_train,
    random_state=42
)
```

Обучающая выборка дополнительно разбивается на train и validation части. Это необходимо для корректной настройки гиперпараметров без утечки данных из тестовой выборки.

Настройка гиперпараметров Logistic Regression с Optuna

Для улучшения baseline-модели используется библиотека Optuna, которая автоматически подбирает оптимальные значения гиперпараметров.

Оптимизируются:

- коэффициент регуляризации `C`
- тип регуляризации (`l1` / `l2`)
- порог классификации

В качестве целевой функции используется произведение F1-score и ROC-AUC, что позволяет учитывать баланс между качеством ранжирования и точностью классификации.

In [20]:
```
def objective(trial):

    # 1️⃣ Гиперпараметры
    C = trial.suggest_float("C", 1e-3, 100, log=True)
    penalty = trial.suggest_categorical("penalty", ["l1", "l2"])
    threshold = trial.suggest_float("threshold", 0.2, 0.8)

    model = LogisticRegression(
        C=C,
        penalty=penalty,
        solver="liblinear",
        class_weight="balanced",
        max_iter=3000,
        random_state=42
    )

    # 2️⃣ Обучение ТОЛЬКО на train
    model.fit(X_tr, y_tr)

    # 3️⃣ Предсказания ТОЛЬКО на validation
    y_proba = model.predict_proba(X_val)[:, 1]

    # 4️⃣ Метрики
    roc_auc = roc_auc_score(y_val, y_proba)
```

```
    y_pred = (y_proba >= threshold).astype(int)
    f1 = f1_score(y_val, y_pred)

    # 5️⃣ Целевая функция
    return f1 * roc_auc
```

In [21]:
```
study = optuna.create_study(
    direction="maximize",
    study_name="LogReg_F1_ROCAUC_Tuning"
)

study.optimize(objective, n_trials=50)
```

```
[I 2025-12-19 23:06:41,150] A new study created in memory with name: LogReg_F1_ROCA
UC_Tuning
[I 2025-12-19 23:06:41,156] Trial 0 finished with value: 0.0 and parameters: {'C':
0.012398415564761251, 'penalty': 'l1', 'threshold': 0.6428800426841497}. Best is tr
ial 0 with value: 0.0.
[I 2025-12-19 23:06:41,162] Trial 1 finished with value: 0.24040574809805582 and pa
rameters: {'C': 0.04246046724020609, 'penalty': 'l1', 'threshold': 0.67095430770246
89}. Best is trial 1 with value: 0.24040574809805582.
[I 2025-12-19 23:06:41,156] Trial 0 finished with value: 0.0 and parameters: {'C':
0.012398415564761251, 'penalty': 'l1', 'threshold': 0.6428800426841497}. Best is tr
ial 0 with value: 0.0.
[I 2025-12-19 23:06:41,162] Trial 1 finished with value: 0.24040574809805582 and pa
rameters: {'C': 0.04246046724020609, 'penalty': 'l1', 'threshold': 0.67095430770246
89}. Best is trial 1 with value: 0.24040574809805582.
[I 2025-12-19 23:06:41,171] Trial 2 finished with value: 0.5715117488043252 and par
ameters: {'C': 0.2280001353710168, 'penalty': 'l1', 'threshold': 0.631401234415808
2}. Best is trial 2 with value: 0.5715117488043252.
[I 2025-12-19 23:06:41,229] Trial 3 finished with value: 0.6588742236024844 and par
ameters: {'C': 13.839971545177223, 'penalty': 'l1', 'threshold': 0.419538041349618
3}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,237] Trial 4 finished with value: 0.6297933513027854 and par
ameters: {'C': 19.28104302547023, 'penalty': 'l2', 'threshold': 0.494606776264468
4}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,242] Trial 5 finished with value: 0.253469387755102 and para
meters: {'C': 0.006706688731216455, 'penalty': 'l2', 'threshold': 0.669353511302160
6}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,247] Trial 6 finished with value: 0.4117647058823529 and par
ameters: {'C': 0.0035952551554759237, 'penalty': 'l1', 'threshold': 0.3313282023285
7774}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,253] Trial 7 finished with value: 0.5031595576619274 and par
ameters: {'C': 0.1632404107143123, 'penalty': 'l1', 'threshold': 0.686129493999413
1}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,262] Trial 8 finished with value: 0.333489417989418 and para
meters: {'C': 0.08467490993353019, 'penalty': 'l1', 'threshold': 0.720337047443295
5}. Best is trial 3 with value: 0.6588742236024844.
[I 2025-12-19 23:06:41,269] Trial 9 finished with value: 0.7004185623293903 and par
ameters: {'C': 1.3548161654357134, 'penalty': 'l2', 'threshold': 0.2279878887172657
5}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,281] Trial 10 finished with value: 0.6939879013494649 and pa
rameters: {'C': 1.592580079415946, 'penalty': 'l2', 'threshold': 0.2501017933033426
5}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,292] Trial 11 finished with value: 0.6919513294276701 and pa
rameters: {'C': 1.8299307519515575, 'penalty': 'l2', 'threshold': 0.208609096577849
1}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,304] Trial 12 finished with value: 0.6955652890086853 and pa
rameters: {'C': 1.5532905865423723, 'penalty': 'l2', 'threshold': 0.206155862541393
5}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,316] Trial 13 finished with value: 0.6862447359896339 and pa
rameters: {'C': 1.5998421067626294, 'penalty': 'l2', 'threshold': 0.330220470156524
8}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,329] Trial 14 finished with value: 0.684831081081081 and par
ameters: {'C': 4.815816787598688, 'penalty': 'l2', 'threshold': 0.3014645034424338
6}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,345] Trial 15 finished with value: 0.6564371980676328 and pa
rameters: {'C': 78.33629116204756, 'penalty': 'l2', 'threshold': 0.422399209151302
4}. Best is trial 9 with value: 0.7004185623293903.
[I 2025-12-19 23:06:41,358] Trial 16 finished with value: 0.7009180474697716 and pa
rameters: {'C': 0.6671965651916241, 'penalty': 'l2', 'threshold': 0.226447386350494
97}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,369] Trial 17 finished with value: 0.626764705882353 and par
```

```
ameters: {'C': 0.0010425847135371918, 'penalty': 'l2', 'threshold': 0.3847771607181
2595}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,381] Trial 18 finished with value: 0.6405985319028797 and pa
rameters: {'C': 0.4971026106203179, 'penalty': 'l2', 'threshold': 0.568006578470312
6}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,392] Trial 19 finished with value: 0.6870045143439638 and pa
rameters: {'C': 0.02434196206712649, 'penalty': 'l2', 'threshold': 0.27491548534421
36}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,404] Trial 20 finished with value: 0.49941886707289557 and p
arameters: {'C': 0.4692232584481654, 'penalty': 'l2', 'threshold': 0.77312128422887
02}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,417] Trial 21 finished with value: 0.6915832575068243 and pa
rameters: {'C': 5.745802359568561, 'penalty': 'l2', 'threshold': 0.2105565618641958
3}. Best is trial 16 with value: 0.7009180474697716.
[I 2025-12-19 23:06:41,430] Trial 22 finished with value: 0.703587403012323 and par
ameters: {'C': 0.8233467298599467, 'penalty': 'l2', 'threshold': 0.2408305082359091
8}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,441] Trial 23 finished with value: 0.6732964453386988 and pa
rameters: {'C': 0.5458874894470972, 'penalty': 'l2', 'threshold': 0.374678315725000
6}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,454] Trial 24 finished with value: 0.6955579129341506 and pa
rameters: {'C': 4.440660388189099, 'penalty': 'l2', 'threshold': 0.262613237062549
7}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,466] Trial 25 finished with value: 0.6387445887445886 and pa
rameters: {'C': 0.0815370228332383, 'penalty': 'l2', 'threshold': 0.495100007997843
9}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,477] Trial 26 finished with value: 0.6949654118404118 and pa
rameters: {'C': 0.6234176433305866, 'penalty': 'l2', 'threshold': 0.316930692470093
76}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,489] Trial 27 finished with value: 0.6968045112781955 and pa
rameters: {'C': 19.792795297897314, 'penalty': 'l2', 'threshold': 0.245074028973944
8}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,500] Trial 28 finished with value: 0.6420094191522763 and pa
rameters: {'C': 0.15140376831695512, 'penalty': 'l2', 'threshold': 0.443451744765814
894}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,512] Trial 29 finished with value: 0.6734615384615384 and pa
rameters: {'C': 0.9158241297407023, 'penalty': 'l2', 'threshold': 0.371350650702559
9}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,525] Trial 30 finished with value: 0.6915833333333333 and pa
rameters: {'C': 8.654855466607128, 'penalty': 'l2', 'threshold': 0.277915250286038
1}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,537] Trial 31 finished with value: 0.6939016393442624 and pa
rameters: {'C': 35.28213909478536, 'penalty': 'l2', 'threshold': 0.2424803550777730
2}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,548] Trial 32 finished with value: 0.6874618585298197 and pa
rameters: {'C': 3.381549849949197, 'penalty': 'l2', 'threshold': 0.2332654878532485
2}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,561] Trial 33 finished with value: 0.6863768115942028 and pa
rameters: {'C': 95.79993991843953, 'penalty': 'l2', 'threshold': 0.285643456018000
1}. Best is trial 22 with value: 0.703587403012323.
[I 2025-12-19 23:06:41,572] Trial 34 finished with value: 0.7067357142857142 and pa
rameters: {'C': 0.27781224307242885, 'penalty': 'l1', 'threshold': 0.342957822120729
97}. Best is trial 34 with value: 0.7067357142857142.
[I 2025-12-19 23:06:41,584] Trial 35 finished with value: 0.7117881112176413 and pa
rameters: {'C': 0.2897981430194464, 'penalty': 'l1', 'threshold': 0.351879977277173
8}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,598] Trial 36 finished with value: 0.63544109277177 and para
meters: {'C': 0.2943232514871735, 'penalty': 'l1', 'threshold': 0.561664092179901
1}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,608] Trial 37 finished with value: 0.6421689101172383 and pa
```

```
rameters: {'C': 0.04856529223185413, 'penalty': 'l1', 'threshold': 0.33736598056830
336}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,620] Trial 38 finished with value: 0.675621156211562 and par
ameters: {'C': 0.2257129461447987, 'penalty': 'l1', 'threshold': 0.4586415251561311
6}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,630] Trial 39 finished with value: 0.45352941176470585 and p
arameters: {'C': 0.017421549221216498, 'penalty': 'l1', 'threshold': 0.369413925498
9512}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,641] Trial 40 finished with value: 0.6999862792574657 and pa
rameters: {'C': 0.09827192336182934, 'penalty': 'l1', 'threshold': 0.40722771716342
077}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,660] Trial 41 finished with value: 0.700695238095238 and par
ameters: {'C': 0.9978609068821767, 'penalty': 'l1', 'threshold': 0.3027764737691438
5}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,677] Trial 42 finished with value: 0.7078792179123304 and pa
rameters: {'C': 0.9069322352968476, 'penalty': 'l1', 'threshold': 0.306150456052388
9}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,700] Trial 43 finished with value: 0.6764755313890262 and pa
rameters: {'C': 2.5044519481127403, 'penalty': 'l1', 'threshold': 0.350497536973722
9}. Best is trial 35 with value: 0.7117881112176413.
[I 2025-12-19 23:06:41,714] Trial 44 finished with value: 0.7193134535367545 and pa
rameters: {'C': 0.3110043457531729, 'penalty': 'l1', 'threshold': 0.302438744036497
56}. Best is trial 44 with value: 0.7193134535367545.
[I 2025-12-19 23:06:41,729] Trial 45 finished with value: 0.716199752628324 and par
ameters: {'C': 0.34018339259220876, 'penalty': 'l1', 'threshold': 0.304903290207263
14}. Best is trial 44 with value: 0.7193134535367545.
[I 2025-12-19 23:06:41,743] Trial 46 finished with value: 0.7024357142857143 and pa
rameters: {'C': 0.18558798389164088, 'penalty': 'l1', 'threshold': 0.34901021110037
4}. Best is trial 44 with value: 0.7193134535367545.
[I 2025-12-19 23:06:41,756] Trial 47 finished with value: 0.6094171259008053 and pa
rameters: {'C': 0.0401928573981733, 'penalty': 'l1', 'threshold': 0.30281038691074
9}. Best is trial 44 with value: 0.7193134535367545.
[I 2025-12-19 23:06:41,774] Trial 48 finished with value: 0.6867945482079758 and pa
rameters: {'C': 0.3159754147393498, 'penalty': 'l1', 'threshold': 0.405801937238195
46}. Best is trial 44 with value: 0.7193134535367545.
[I 2025-12-19 23:06:41,788] Trial 49 finished with value: 0.6826173826173826 and pa
rameters: {'C': 0.09931807849166928, 'penalty': 'l1', 'threshold': 0.43835714261718
56}. Best is trial 44 with value: 0.7193134535367545.
```

In [22]:
```
study.best_params
```

Out[22]:
```
{'C': 0.3110043457531729, 'penalty': 'l1', 'threshold': 0.30243874403649756}
```

Процедура оптимизации успешно завершена. Найдена комбинация гиперпараметров, превосходящая baseline-конфигурацию по выбранной целевой метрике.

Обучение и оценка оптимизированной Logistic Regression

На данном этапе модель логистической регрессии обучается с использованием найденных оптимальных гиперпараметров.

Оценка проводится на тестовой выборке, которая не участвовала ни в обучении, ни в подборе параметров.

In [23]:
```
best_params = study.best_params

lr_tuned = LogisticRegression(
    C=best_params["C"],
    penalty=best_params["penalty"],
    solver="liblinear",
    class_weight="balanced",
    max_iter=3000,
    random_state=42
)

lr_tuned.fit(X_train_lr, y_train)
```

Out[23]:

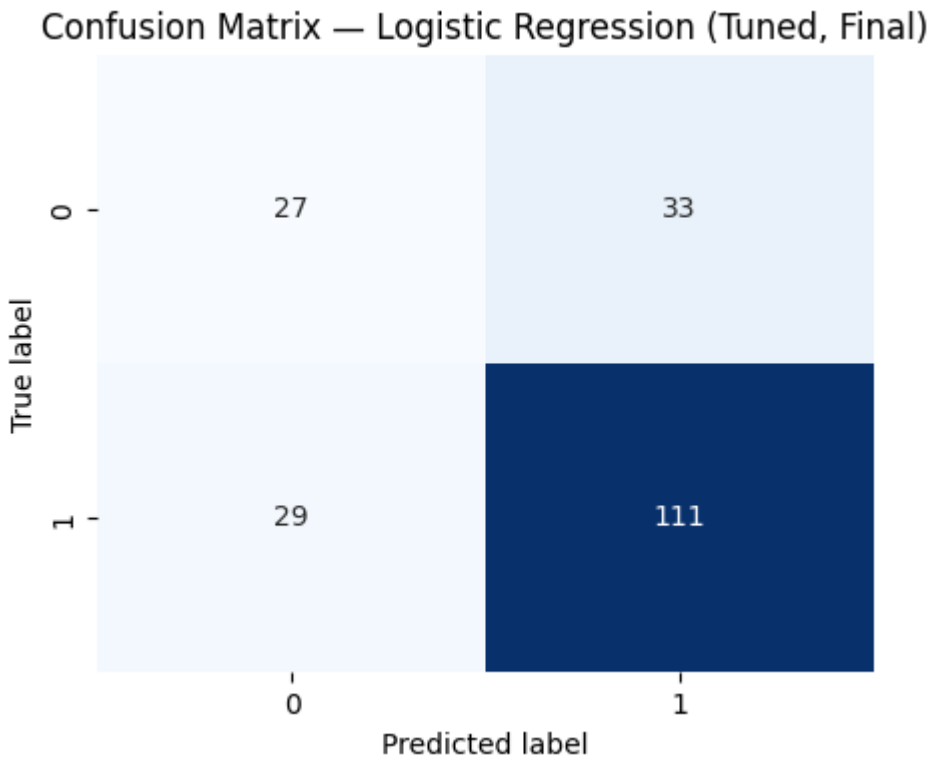
▾ LogisticRegression

▶ Parameters

In [24]:
```
y_proba_test = lr_tuned.predict_proba(X_test_lr)[:, 1]

evaluate_classification_model(
    y_true=y_test,
    y_proba=y_proba_test,
    threshold=best_params["threshold"],
    model_name="Logistic Regression (Tuned, Final)"
)
```

```
Logistic Regression (Tuned, Final)
----------------------------------
ROC-AUC : 0.7592
F1-score: 0.7817
Threshold: 0.30243874403649756
```



Out[24]:
```
{'model': 'Logistic Regression (Tuned, Final)',
 'roc_auc': 0.7591666666666668,
 'f1_score': 0.7816901408450704,
 'threshold': 0.30243874403649756}
```

Оптимизированная модель демонстрирует улучшение F1-score по сравнению с baseline-моделью при сопоставимом значении ROC-AUC.

Это подтверждает эффективность подбора гиперпараметров и оптимизации порога.

In [26]:
```
os.makedirs("models", exist_ok=True)

lr_tuned_artifact = {
    "model": lr_tuned,
    "preprocessor": preprocessor,
    "best_params": {
        "C": best_params["C"],
        "penalty": best_params["penalty"],
        "threshold": best_params["threshold"]
    },
    "metrics": {
        "roc_auc": 0.7592,
        "f1_score": 0.7817
    }
}

joblib.dump(
    lr_tuned_artifact,
    "models/lr_tuned_model.joblib"
)
```

Out[26]:
```
['models/lr_tuned_model.joblib']
```

Настройка и оценка Random Forest

После логистической регрессии рассматривается более сложная модель — Random Forest, способная учитывать нелинейные зависимости между признаками.

С помощью Optuna подбираются:

- параметры структуры деревьев
- количество деревьев
- порог классификации

In [27]:
```
rf_data = joblib.load("rf_preprocessing.joblib")

X_train_rf = rf_data["X_train"]
X_test_rf  = rf_data["X_test"]
y_train    = rf_data["y_train"]
y_test     = rf_data["y_test"]

rf_preprocessor = rf_data["preprocessor"]
```

In [28]:
```
X_train_rf_enc = rf_preprocessor.fit_transform(X_train_rf)
X_test_rf_enc  = rf_preprocessor.transform(X_test_rf)
```

```
X_train_rf_enc.shape, X_test_rf_enc.shape
```

Out[28]: ((800, 65), (200, 65))

In [29]:
```
X_tr, X_val, y_tr, y_val = train_test_split(
    X_train_rf_enc,
    y_train,
    test_size=0.25,
    stratify=y_train,
    random_state=42
)
```

In [31]:
```
def objective(trial):

    # 1️⃣ Гиперпараметры Random Forest
    n_estimators = trial.suggest_int("n_estimators", 200, 800)
    max_depth = trial.suggest_int("max_depth", 4, 20)
    min_samples_split = trial.suggest_int("min_samples_split", 2, 20)
    min_samples_leaf = trial.suggest_int("min_samples_leaf", 1, 15)
    max_features = trial.suggest_categorical(
        "max_features", ["sqrt", "log2", None]
    )

    # threshold — тоже гиперпараметр
    threshold = trial.suggest_float("threshold", 0.2, 0.8)

    model = RandomForestClassifier(
        n_estimators=n_estimators,
        max_depth=max_depth,
        min_samples_split=min_samples_split,
        min_samples_leaf=min_samples_leaf,
        max_features=max_features,
        class_weight="balanced",
        random_state=42,
        n_jobs=-1
    )

    # 2️⃣ Обучение
    model.fit(X_tr, y_tr)

    # 3️⃣ Вероятности на validation
    y_proba = model.predict_proba(X_val)[:, 1]

    # 4️⃣ Метрики
    roc_auc = roc_auc_score(y_val, y_proba)
    y_pred = (y_proba >= threshold).astype(int)
    f1 = f1_score(y_val, y_pred)

    # 5️⃣ Целевая функция
    return f1 * roc_auc
```

In [32]:
```
study_rf = optuna.create_study(
    direction="maximize",
    study_name="RandomForest_F1_ROCAUC_Tuning"
)

study_rf.optimize(objective, n_trials=50)
```

```
[I 2025-12-19 23:12:17,444] A new study created in memory with name: RandomForest_F
1_ROCAUC_Tuning
[I 2025-12-19 23:12:18,516] Trial 0 finished with value: 0.6674509803921568 and par
ameters: {'n_estimators': 729, 'max_depth': 16, 'min_samples_split': 20, 'min_sampl
es_leaf': 8, 'max_features': 'log2', 'threshold': 0.21465043694525537}. Best is tri
al 0 with value: 0.6674509803921568.
[I 2025-12-19 23:12:19,768] Trial 1 finished with value: 0.6924471299093655 and par
ameters: {'n_estimators': 732, 'max_depth': 6, 'min_samples_split': 3, 'min_samples
_leaf': 9, 'max_features': 'sqrt', 'threshold': 0.3348470521684109}. Best is trial
1 with value: 0.6924471299093655.
[I 2025-12-19 23:12:20,176] Trial 2 finished with value: 0.4553317535545023 and par
ameters: {'n_estimators': 215, 'max_depth': 20, 'min_samples_split': 2, 'min_sample
s_leaf': 15, 'max_features': None, 'threshold': 0.7352717559021007}. Best is trial
1 with value: 0.6924471299093655.
[I 2025-12-19 23:12:21,225] Trial 3 finished with value: 0.6988500663423264 and par
ameters: {'n_estimators': 654, 'max_depth': 6, 'min_samples_split': 2, 'min_samples
_leaf': 11, 'max_features': 'log2', 'threshold': 0.37570181029288874}. Best is tria
l 3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:21,971] Trial 4 finished with value: 0.6262910798122066 and par
ameters: {'n_estimators': 402, 'max_depth': 16, 'min_samples_split': 6, 'min_sample
s_leaf': 13, 'max_features': None, 'threshold': 0.4770123057511457}. Best is trial
3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:22,730] Trial 5 finished with value: 0.27923669467787116 and pa
rameters: {'n_estimators': 458, 'max_depth': 7, 'min_samples_split': 4, 'min_sample
s_leaf': 8, 'max_features': 'log2', 'threshold': 0.7586699477423287}. Best is trial
3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:24,088] Trial 6 finished with value: 0.497992277992278 and para
meters: {'n_estimators': 726, 'max_depth': 18, 'min_samples_split': 18, 'min_sample
s_leaf': 14, 'max_features': None, 'threshold': 0.7127078787965744}. Best is trial
3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:24,758] Trial 7 finished with value: 0.6687254901960784 and par
ameters: {'n_estimators': 399, 'max_depth': 15, 'min_samples_split': 19, 'min_sampl
es_leaf': 2, 'max_features': 'log2', 'threshold': 0.21938343251094075}. Best is tri
al 3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:25,192] Trial 8 finished with value: 0.5077417989417989 and par
ameters: {'n_estimators': 216, 'max_depth': 7, 'min_samples_split': 13, 'min_sample
s_leaf': 10, 'max_features': None, 'threshold': 0.7217151716882275}. Best is trial
3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:26,458] Trial 9 finished with value: 0.6653883972468043 and par
ameters: {'n_estimators': 725, 'max_depth': 18, 'min_samples_split': 8, 'min_sample
s_leaf': 1, 'max_features': 'log2', 'threshold': 0.31775396640131565}. Best is tria
l 3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:27,485] Trial 10 finished with value: 0.6754770318021202 and pa
rameters: {'n_estimators': 562, 'max_depth': 11, 'min_samples_split': 12, 'min_samp
les_leaf': 4, 'max_features': 'sqrt', 'threshold': 0.5422732565918947}. Best is tri
al 3 with value: 0.6988500663423264.
[I 2025-12-19 23:12:28,437] Trial 11 finished with value: 0.6999382716049383 and pa
rameters: {'n_estimators': 608, 'max_depth': 4, 'min_samples_split': 2, 'min_sample
s_leaf': 11, 'max_features': 'sqrt', 'threshold': 0.38998632928634924}. Best is tri
al 11 with value: 0.6999382716049383.
[I 2025-12-19 23:12:29,402] Trial 12 finished with value: 0.701259987216363 and par
ameters: {'n_estimators': 590, 'max_depth': 4, 'min_samples_split': 8, 'min_samples
_leaf': 11, 'max_features': 'sqrt', 'threshold': 0.46343987818454424}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:30,346] Trial 13 finished with value: 0.6652453653217012 and pa
rameters: {'n_estimators': 571, 'max_depth': 4, 'min_samples_split': 9, 'min_sample
s_leaf': 12, 'max_features': 'sqrt', 'threshold': 0.5218405055675335}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:31,453] Trial 14 finished with value: 0.698076923076923 and par
ameters: {'n_estimators': 604, 'max_depth': 10, 'min_samples_split': 15, 'min_sampl
```

```
es_leaf': 5, 'max_features': 'sqrt', 'threshold': 0.4198088136923834}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:32,433] Trial 15 finished with value: 0.46552188552188556 and p
arameters: {'n_estimators': 489, 'max_depth': 4, 'min_samples_split': 6, 'min_sampl
es_leaf': 11, 'max_features': 'sqrt', 'threshold': 0.625240338406199}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:33,503] Trial 16 finished with value: 0.5949941451990632 and pa
rameters: {'n_estimators': 642, 'max_depth': 10, 'min_samples_split': 9, 'min_sampl
es_leaf': 6, 'max_features': 'sqrt', 'threshold': 0.6045631290004272}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:34,052] Trial 17 finished with value: 0.6829124423963133 and pa
rameters: {'n_estimators': 328, 'max_depth': 13, 'min_samples_split': 5, 'min_sampl
es_leaf': 7, 'max_features': 'sqrt', 'threshold': 0.43509437511715}. Best is trial
12 with value: 0.701259987216363.
[I 2025-12-19 23:12:35,356] Trial 18 finished with value: 0.6678466076696165 and pa
rameters: {'n_estimators': 786, 'max_depth': 9, 'min_samples_split': 15, 'min_sampl
es_leaf': 13, 'max_features': 'sqrt', 'threshold': 0.3015245578922825}. Best is tri
al 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:36,246] Trial 19 finished with value: 0.49256268691051297 and p
arameters: {'n_estimators': 523, 'max_depth': 4, 'min_samples_split': 7, 'min_sampl
es_leaf': 10, 'max_features': 'sqrt', 'threshold': 0.613186099624317}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:37,364] Trial 20 finished with value: 0.6794465894465894 and pa
rameters: {'n_estimators': 652, 'max_depth': 13, 'min_samples_split': 10, 'min_samp
les_leaf': 15, 'max_features': 'sqrt', 'threshold': 0.4628105484566905}. Best is tr
ial 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:38,487] Trial 21 finished with value: 0.6951785714285714 and pa
rameters: {'n_estimators': 654, 'max_depth': 6, 'min_samples_split': 2, 'min_sample
s_leaf': 11, 'max_features': 'log2', 'threshold': 0.3926821943588483}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:39,466] Trial 22 finished with value: 0.6946138211382114 and pa
rameters: {'n_estimators': 599, 'max_depth': 6, 'min_samples_split': 4, 'min_sample
s_leaf': 12, 'max_features': 'log2', 'threshold': 0.3629231738692588}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:40,531] Trial 23 finished with value: 0.6947249334516415 and pa
rameters: {'n_estimators': 678, 'max_depth': 8, 'min_samples_split': 2, 'min_sample
s_leaf': 10, 'max_features': 'log2', 'threshold': 0.38200765229820377}. Best is tri
al 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:41,406] Trial 24 finished with value: 0.6709803921568628 and pa
rameters: {'n_estimators': 540, 'max_depth': 5, 'min_samples_split': 4, 'min_sample
s_leaf': 12, 'max_features': 'sqrt', 'threshold': 0.2723094201150013}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:42,376] Trial 25 finished with value: 0.6246101309049517 and pa
rameters: {'n_estimators': 606, 'max_depth': 8, 'min_samples_split': 6, 'min_sample
s_leaf': 11, 'max_features': 'log2', 'threshold': 0.5565966752266949}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:43,636] Trial 26 finished with value: 0.6768627450980391 and pa
rameters: {'n_estimators': 800, 'max_depth': 5, 'min_samples_split': 11, 'min_sampl
es_leaf': 9, 'max_features': 'sqrt', 'threshold': 0.26586550042416035}. Best is tri
al 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:44,469] Trial 27 finished with value: 0.6974027388733272 and pa
rameters: {'n_estimators': 473, 'max_depth': 5, 'min_samples_split': 4, 'min_sample
s_leaf': 13, 'max_features': 'sqrt', 'threshold': 0.43325409626146993}. Best is tri
al 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:45,717] Trial 28 finished with value: 0.6574121405750799 and pa
rameters: {'n_estimators': 686, 'max_depth': 8, 'min_samples_split': 8, 'min_sample
s_leaf': 9, 'max_features': None, 'threshold': 0.3528937408602327}. Best is trial 1
2 with value: 0.701259987216363.
[I 2025-12-19 23:12:46,451] Trial 29 finished with value: 0.6882411674347158 and pa
rameters: {'n_estimators': 423, 'max_depth': 4, 'min_samples_split': 3, 'min_sample
```

```
s_leaf': 8, 'max_features': 'log2', 'threshold': 0.49156449653940637}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:47,525] Trial 30 finished with value: 0.6956849932401983 and pa
rameters: {'n_estimators': 605, 'max_depth': 7, 'min_samples_split': 16, 'min_sampl
es_leaf': 14, 'max_features': 'log2', 'threshold': 0.404767943876716}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:48,653] Trial 31 finished with value: 0.7000060661207159 and pa
rameters: {'n_estimators': 615, 'max_depth': 10, 'min_samples_split': 14, 'min_samp
les_leaf': 5, 'max_features': 'sqrt', 'threshold': 0.431088738022838}. Best is tria
l 12 with value: 0.701259987216363.
[I 2025-12-19 23:12:49,789] Trial 32 finished with value: 0.7042998585572843 and pa
rameters: {'n_estimators': 697, 'max_depth': 6, 'min_samples_split': 13, 'min_sampl
es_leaf': 3, 'max_features': 'sqrt', 'threshold': 0.44966634595823485}. Best is tri
al 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:50,911] Trial 33 finished with value: 0.691303669008587 and par
ameters: {'n_estimators': 701, 'max_depth': 12, 'min_samples_split': 13, 'min_sampl
es_leaf': 3, 'max_features': 'sqrt', 'threshold': 0.46196820801941013}. Best is tri
al 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:51,752] Trial 34 finished with value: 0.595874349739896 and par
ameters: {'n_estimators': 523, 'max_depth': 5, 'min_samples_split': 14, 'min_sample
s_leaf': 5, 'max_features': 'sqrt', 'threshold': 0.5735696018690198}. Best is trial
32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:52,772] Trial 35 finished with value: 0.69 and parameters: {'n_
estimators': 631, 'max_depth': 9, 'min_samples_split': 11, 'min_samples_leaf': 3,
'max_features': 'sqrt', 'threshold': 0.5022345497875799}. Best is trial 32 with val
ue: 0.7042998585572843.
[I 2025-12-19 23:12:53,965] Trial 36 finished with value: 0.692000628634292 and par
ameters: {'n_estimators': 766, 'max_depth': 6, 'min_samples_split': 16, 'min_sample
s_leaf': 7, 'max_features': 'sqrt', 'threshold': 0.45250726899771404}. Best is tria
l 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:54,881] Trial 37 finished with value: 0.7010540069686412 and pa
rameters: {'n_estimators': 573, 'max_depth': 7, 'min_samples_split': 12, 'min_sampl
es_leaf': 1, 'max_features': 'sqrt', 'threshold': 0.3462169810849901}. Best is tria
l 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:55,794] Trial 38 finished with value: 0.6885815295815296 and pa
rameters: {'n_estimators': 566, 'max_depth': 9, 'min_samples_split': 12, 'min_sampl
es_leaf': 1, 'max_features': 'sqrt', 'threshold': 0.3382104455532876}. Best is tria
l 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:57,258] Trial 39 finished with value: 0.6685000829600134 and pa
rameters: {'n_estimators': 760, 'max_depth': 7, 'min_samples_split': 17, 'min_sampl
es_leaf': 2, 'max_features': None, 'threshold': 0.5181242645734019}. Best is trial
32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:58,383] Trial 40 finished with value: 0.6680392156862746 and pa
rameters: {'n_estimators': 706, 'max_depth': 11, 'min_samples_split': 20, 'min_samp
les_leaf': 3, 'max_features': 'sqrt', 'threshold': 0.20706909400576506}. Best is tr
ial 32 with value: 0.7042998585572843.
[I 2025-12-19 23:12:59,287] Trial 41 finished with value: 0.7138961038961039 and pa
rameters: {'n_estimators': 566, 'max_depth': 6, 'min_samples_split': 13, 'min_sampl
es_leaf': 2, 'max_features': 'sqrt', 'threshold': 0.40444391169605093}. Best is tri
al 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:00,163] Trial 42 finished with value: 0.6849653808110782 and pa
rameters: {'n_estimators': 548, 'max_depth': 6, 'min_samples_split': 13, 'min_sampl
es_leaf': 2, 'max_features': 'sqrt', 'threshold': 0.30909988228604224}. Best is tri
al 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:01,082] Trial 43 finished with value: 0.6958585858585858 and pa
rameters: {'n_estimators': 575, 'max_depth': 7, 'min_samples_split': 14, 'min_sampl
es_leaf': 4, 'max_features': 'sqrt', 'threshold': 0.4801559285772759}. Best is tria
l 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:01,932] Trial 44 finished with value: 0.708805031446541 and par
ameters: {'n_estimators': 524, 'max_depth': 8, 'min_samples_split': 11, 'min_sample
```

```
s_leaf': 1, 'max_features': 'sqrt', 'threshold': 0.41347223827872}. Best is trial 4
1 with value: 0.7138961038961039.
[I 2025-12-19 23:13:02,744] Trial 45 finished with value: 0.6605921052631579 and pa
rameters: {'n_estimators': 449, 'max_depth': 5, 'min_samples_split': 11, 'min_sampl
es_leaf': 1, 'max_features': None, 'threshold': 0.4122143016193234}. Best is trial
41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:03,402] Trial 46 finished with value: 0.7002628513178971 and pa
rameters: {'n_estimators': 346, 'max_depth': 8, 'min_samples_split': 10, 'min_sampl
es_leaf': 1, 'max_features': 'sqrt', 'threshold': 0.35465933827548657}. Best is tri
al 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:04,332] Trial 47 finished with value: 0.6824975417895772 and pa
rameters: {'n_estimators': 506, 'max_depth': 7, 'min_samples_split': 12, 'min_sampl
es_leaf': 2, 'max_features': 'sqrt', 'threshold': 0.27373367920541936}. Best is tri
al 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:05,176] Trial 48 finished with value: 0.4221989528795811 and pa
rameters: {'n_estimators': 507, 'max_depth': 5, 'min_samples_split': 10, 'min_sampl
es_leaf': 4, 'max_features': 'sqrt', 'threshold': 0.6811593415972107}. Best is tria
l 41 with value: 0.7138961038961039.
[I 2025-12-19 23:13:06,118] Trial 49 finished with value: 0.3816164736164736 and pa
rameters: {'n_estimators': 578, 'max_depth': 19, 'min_samples_split': 12, 'min_samp
les_leaf': 2, 'max_features': 'sqrt', 'threshold': 0.7954883997293497}. Best is tri
al 41 with value: 0.7138961038961039.
```

In [33]:
```
study_rf.best_params
```

Out[33]:
```
{'n_estimators': 566,
 'max_depth': 6,
 'min_samples_split': 13,
 'min_samples_leaf': 2,
 'max_features': 'sqrt',
 'threshold': 0.40444391169605093}
```

Наилучшие гиперпараметры были успешно подобраны

Обучаем модель с наилучшими гиперпараметрами Random Forest

In [34]:
```
best_params = study_rf.best_params

rf_tuned = RandomForestClassifier(
    n_estimators=best_params["n_estimators"],
    max_depth=best_params["max_depth"],
    min_samples_split=best_params["min_samples_split"],
    min_samples_leaf=best_params["min_samples_leaf"],
    max_features=best_params["max_features"],
    class_weight="balanced",
    random_state=42,
    n_jobs=-1
)

rf_tuned.fit(X_train_rf_enc, y_train)
```

Out[34]:

▾ RandomForestClassifier

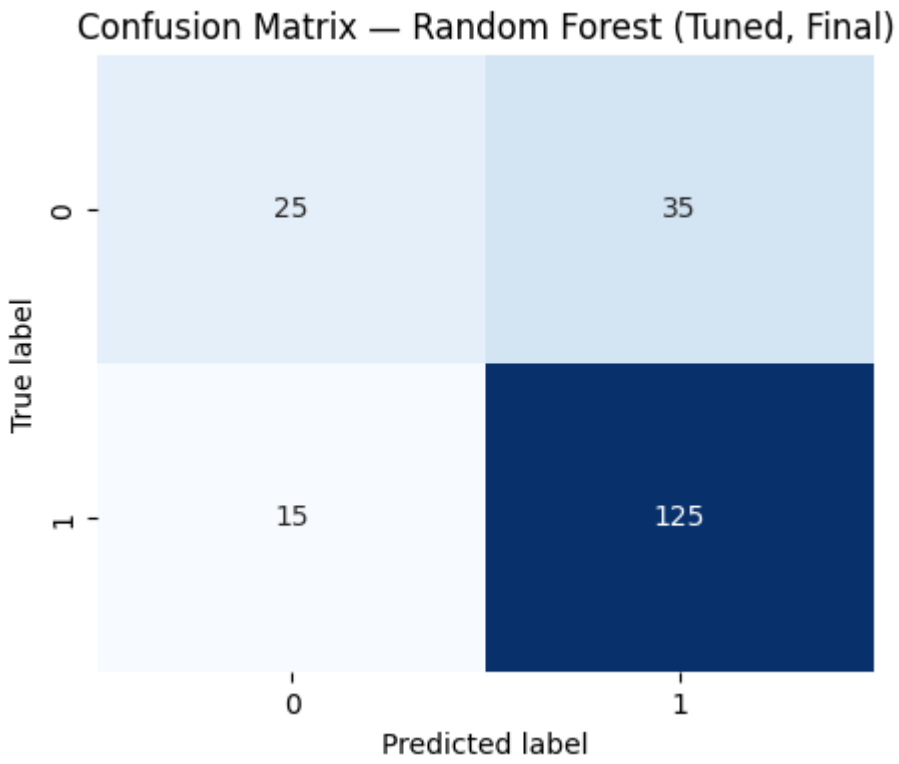
▶ Parameters

Оценки метрики качества

In [35]:
```
y_proba_rf_test = rf_tuned.predict_proba(X_test_rf_enc)[:, 1]

evaluate_classification_model(
    y_true=y_test,
    y_proba=y_proba_rf_test,
    threshold=best_params["threshold"],
    model_name="Random Forest (Tuned, Final)"
)
```

```
Random Forest (Tuned, Final)
----------------------------
ROC-AUC : 0.7881
F1-score: 0.8333
Threshold: 0.40444391169605093
```



Out[35]:
```
{'model': 'Random Forest (Tuned, Final)',
 'roc_auc': 0.7880952380952381,
 'f1_score': 0.8333333333333334,
 'threshold': 0.40444391169605093}
```

Оптимизированная модель Random Forest показывает заметный прирост как по ROC-AUC, так и по F1-score, превосходя логистическую регрессию.

Настройка и оценка CatBoost

В качестве финальной модели используется CatBoost, который эффективно работает с категориальными признаками и часто показывает высокое качество на табличных данных.

Оптимизация включает подбор:

- глубины деревьев
- скорости обучения
- регуляризации
- порога классификации

In [36]:
```
cb_data = joblib.load("catboost_preprocessing.joblib")

X_train_cb = cb_data["X_train"]
X_test_cb  = cb_data["X_test"]
y_train    = cb_data["y_train"]
y_test     = cb_data["y_test"]

cat_feature_indices = cb_data["cat_feature_indices"]
```

In [37]:
```
X_tr, X_val, y_tr, y_val = train_test_split(
    X_train_cb,
    y_train,
    test_size=0.25,
    stratify=y_train,
    random_state=42
)
```

In [39]:
```
def objective(trial):

    params = {
        "iterations": trial.suggest_int("iterations", 400, 1200),
        "depth": trial.suggest_int("depth", 4, 10),
        "learning_rate": trial.suggest_float("learning_rate", 0.01, 0.3, log=True)
        "l2_leaf_reg": trial.suggest_float("l2_leaf_reg", 1, 10),
        "random_strength": trial.suggest_float("random_strength", 0, 1),
        "loss_function": "Logloss",
        "eval_metric": "AUC",
        "verbose": 0,
        "random_state": 42
    }

    threshold = trial.suggest_float("threshold", 0.2, 0.8)

    model = CatBoostClassifier(**params)

    model.fit(
        X_tr, y_tr,
        cat_features=cat_feature_indices,
        eval_set=(X_val, y_val),
        verbose=False
    )

    y_proba = model.predict_proba(X_val)[:, 1]

    roc_auc = roc_auc_score(y_val, y_proba)
    y_pred = (y_proba >= threshold).astype(int)
    f1 = f1_score(y_val, y_pred)

    return f1 * roc_auc
```

Настройка и подбор наилушие гиперпараметров

In [40]:
```
study_cb = optuna.create_study(
    direction="maximize",
    study_name="CatBoost_F1_ROCAUC_Tuning"
)

study_cb.optimize(objective, n_trials=50)
```

```
[I 2025-12-19 23:17:21,304] A new study created in memory with name: CatBoost_F1_RO
CAUC_Tuning
[I 2025-12-19 23:17:56,609] Trial 0 finished with value: 0.6771025751690405 and par
ameters: {'iterations': 854, 'depth': 6, 'learning_rate': 0.11402592312083465, 'l2_
leaf_reg': 8.00688911947498, 'random_strength': 0.5513938569016931, 'threshold': 0.
338441829457132}. Best is trial 0 with value: 0.6771025751690405.
[I 2025-12-19 23:18:18,452] Trial 1 finished with value: 0.6826772814388914 and par
ameters: {'iterations': 879, 'depth': 4, 'learning_rate': 0.14694932600861493, 'l2_
leaf_reg': 4.541115367878135, 'random_strength': 0.3656349027052319, 'threshold':
0.24541323243821372}. Best is trial 1 with value: 0.6826772814388914.
[I 2025-12-19 23:18:39,779] Trial 2 finished with value: 0.6785801393728222 and par
ameters: {'iterations': 849, 'depth': 4, 'learning_rate': 0.01859588816897441, 'l2_
leaf_reg': 5.455041499587826, 'random_strength': 0.6610852788114667, 'threshold':
0.27653978111064775}. Best is trial 1 with value: 0.6826772814388914.
[I 2025-12-19 23:18:53,242] Trial 3 finished with value: 0.6837510955302365 and par
ameters: {'iterations': 528, 'depth': 4, 'learning_rate': 0.025987116326667264, 'l2
_leaf_reg': 1.2327072176099374, 'random_strength': 0.44594628251767976, 'threshol
d': 0.28686134057144785}. Best is trial 3 with value: 0.6837510955302365.
[I 2025-12-19 23:19:26,566] Trial 4 finished with value: 0.6854994742376446 and par
ameters: {'iterations': 711, 'depth': 7, 'learning_rate': 0.16920242274436795, 'l2_
leaf_reg': 6.371007688160894, 'random_strength': 0.026030535040396874, 'threshold':
0.42131321094606766}. Best is trial 4 with value: 0.6854994742376446.
[I 2025-12-19 23:19:49,830] Trial 5 finished with value: 0.6862216898954704 and par
ameters: {'iterations': 909, 'depth': 4, 'learning_rate': 0.021932695521315514, 'l2
_leaf_reg': 5.996928678768707, 'random_strength': 0.47159825637579167, 'threshold':
0.2756977764208008}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:20:11,454] Trial 6 finished with value: 0.6854507337526206 and par
ameters: {'iterations': 470, 'depth': 7, 'learning_rate': 0.06022683661792617, 'l2_
leaf_reg': 3.4414945284051295, 'random_strength': 0.5378354342643439, 'threshold':
0.4343910073003755}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:20:34,437] Trial 7 finished with value: 0.6850125313283207 and par
ameters: {'iterations': 583, 'depth': 6, 'learning_rate': 0.10195096470657689, 'l2_
leaf_reg': 7.251795952320892, 'random_strength': 0.46038615439988695, 'threshold':
0.7093994586250039}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:21:49,461] Trial 8 finished with value: 0.6571085089773614 and par
ameters: {'iterations': 978, 'depth': 9, 'learning_rate': 0.04915776614889487, 'l2_
leaf_reg': 1.8811781916885308, 'random_strength': 0.07319639284232271, 'threshold':
0.561191935724977}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:22:13,221] Trial 9 finished with value: 0.618933410762679 and para
meters: {'iterations': 728, 'depth': 5, 'learning_rate': 0.18070231309662077, 'l2_l
eaf_reg': 9.553177086938577, 'random_strength': 0.42856622373446007, 'threshold':
0.769356335448159}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:23:58,384] Trial 10 finished with value: 0.6602739726027397 and pa
rameters: {'iterations': 1168, 'depth': 10, 'learning_rate': 0.01079315007597357,
'l2_leaf_reg': 9.471981873318725, 'random_strength': 0.9616318143546734, 'threshol
d': 0.5726727783649974}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:24:37,048] Trial 11 finished with value: 0.6789049145299145 and pa
rameters: {'iterations': 647, 'depth': 8, 'learning_rate': 0.290641808157968, 'l2_l
eaf_reg': 6.781770452770074, 'random_strength': 0.012631940114156415, 'threshold':
0.4255361892060322}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:25:25,065] Trial 12 finished with value: 0.6728556776556778 and pa
rameters: {'iterations': 1027, 'depth': 7, 'learning_rate': 0.038017718966274705,
'l2_leaf_reg': 5.785503872396812, 'random_strength': 0.23565787335594615, 'threshol
d': 0.3795007160639169}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:25:52,960] Trial 13 finished with value: 0.6716962524654833 and pa
rameters: {'iterations': 709, 'depth': 6, 'learning_rate': 0.011751549000359046, 'l
2_leaf_reg': 3.8716614771010787, 'random_strength': 0.7700326403596973, 'threshol
d': 0.49609512374467457}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:26:51,293] Trial 14 finished with value: 0.6608268398268398 and pa
rameters: {'iterations': 1014, 'depth': 8, 'learning_rate': 0.06693200914655043, 'l
```

```
2_leaf_reg': 8.049834567736113, 'random_strength': 0.2211742822675135, 'threshold':
0.20865099254694197}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:27:15,699] Trial 15 finished with value: 0.6836890243902439 and pa
rameters: {'iterations': 760, 'depth': 5, 'learning_rate': 0.024699465034715717, 'l
2_leaf_reg': 6.00005292382753, 'random_strength': 0.17827005760736325, 'threshold':
0.34976891265098636}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:29:07,264] Trial 16 finished with value: 0.6576158940397352 and pa
rameters: {'iterations': 1171, 'depth': 10, 'learning_rate': 0.2910433231504041, 'l
2_leaf_reg': 4.749396034832765, 'random_strength': 0.8837786008908075, 'threshold':
0.49655198365253184}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:30:00,113] Trial 17 finished with value: 0.683412162162162 and par
ameters: {'iterations': 925, 'depth': 8, 'learning_rate': 0.017456142282563845, 'l2
_leaf_reg': 2.843678120866854, 'random_strength': 0.3194611057920436, 'threshold':
0.6306820496204552}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:30:20,274] Trial 18 finished with value: 0.6848598989435002 and pa
rameters: {'iterations': 641, 'depth': 5, 'learning_rate': 0.03717600298205967, 'l2
_leaf_reg': 6.895946969497693, 'random_strength': 0.7211940133347188, 'threshold':
0.42746790147547276}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:30:39,975] Trial 19 finished with value: 0.6774025320097827 and pa
rameters: {'iterations': 413, 'depth': 7, 'learning_rate': 0.09972077552262443, 'l2
_leaf_reg': 8.606491186462982, 'random_strength': 0.13082080127258505, 'threshold':
0.31089452627331493}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:31:58,663] Trial 20 finished with value: 0.6767770034843206 and pa
rameters: {'iterations': 1083, 'depth': 9, 'learning_rate': 0.20688181803178018, 'l
2_leaf_reg': 6.297577405666735, 'random_strength': 0.6095788251687372, 'threshold':
0.20469356758468737}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:32:19,432] Trial 21 finished with value: 0.6785377358490566 and pa
rameters: {'iterations': 414, 'depth': 7, 'learning_rate': 0.06855371045505275, 'l2
_leaf_reg': 3.2308939132402887, 'random_strength': 0.5415014471155514, 'threshold':
0.42200982202836324}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:32:38,586] Trial 22 finished with value: 0.6735217954251187 and pa
rameters: {'iterations': 467, 'depth': 6, 'learning_rate': 0.07557529898685372, 'l2
_leaf_reg': 4.54300061892322, 'random_strength': 0.3105365164177919, 'threshold':
0.38589799840749983}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:33:15,530] Trial 23 finished with value: 0.6742537313432836 and pa
rameters: {'iterations': 797, 'depth': 7, 'learning_rate': 0.03401559462833066, 'l2
_leaf_reg': 5.140926595114122, 'random_strength': 0.7610681310082018, 'threshold':
0.46185236015608777}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:33:53,233] Trial 24 finished with value: 0.6721565315315315 and pa
rameters: {'iterations': 513, 'depth': 9, 'learning_rate': 0.049344353183969736, 'l
2_leaf_reg': 3.618808295380596, 'random_strength': 0.8430782668966911, 'threshold':
0.5827220487237017}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:34:33,121] Trial 25 finished with value: 0.6802827838827838 and pa
rameters: {'iterations': 672, 'depth': 8, 'learning_rate': 0.016334524570655803, 'l
2_leaf_reg': 2.858061406896054, 'random_strength': 0.618666885969431, 'threshold':
0.35694669268284285}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:34:51,442] Trial 26 finished with value: 0.678936877076412 and par
ameters: {'iterations': 586, 'depth': 5, 'learning_rate': 0.1477189273274078, 'l2_l
eaf_reg': 7.429893749957592, 'random_strength': 0.33926589439279764, 'threshold':
0.535233991529211}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:35:32,667] Trial 27 finished with value: 0.6825188834154351 and pa
rameters: {'iterations': 925, 'depth': 7, 'learning_rate': 0.026266292316003723, 'l
2_leaf_reg': 3.9877852998718155, 'random_strength': 0.01102666771398466, 'threshol
d': 0.6400619545808155}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:36:12,627] Trial 28 finished with value: 0.6856497695852535 and pa
rameters: {'iterations': 1098, 'depth': 6, 'learning_rate': 0.07568942558386103, 'l
2_leaf_reg': 2.0068544005155413, 'random_strength': 0.5142763846753268, 'threshol
d': 0.46081186662538504}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:36:53,437] Trial 29 finished with value: 0.6803419913419912 and pa
rameters: {'iterations': 1101, 'depth': 6, 'learning_rate': 0.08295580338657363, 'l
```

```
2_leaf_reg': 8.239038272669466, 'random_strength': 0.2577955538116273, 'threshold':
0.3184024041516811}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:37:18,323] Trial 30 finished with value: 0.660800727934486 and par
ameters: {'iterations': 828, 'depth': 5, 'learning_rate': 0.12730632070640654, 'l2_
leaf_reg': 2.201983415956903, 'random_strength': 0.3975765272790506, 'threshold':
0.2556776576949686}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:37:58,024] Trial 31 finished with value: 0.6843974132863021 and pa
rameters: {'iterations': 1083, 'depth': 6, 'learning_rate': 0.049674739829489604,
'l2_leaf_reg': 1.1742538210510896, 'random_strength': 0.5078925018841434, 'threshol
d': 0.4430015856375965}. Best is trial 5 with value: 0.6862216898954704.
[I 2025-12-19 23:38:20,084] Trial 32 finished with value: 0.6988019891500904 and pa
rameters: {'iterations': 887, 'depth': 4, 'learning_rate': 0.062413307360762345, 'l
2_leaf_reg': 2.0892429304264115, 'random_strength': 0.5715750086644773, 'threshol
d': 0.47484086203061815}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:38:42,929] Trial 33 finished with value: 0.6876142857142857 and pa
rameters: {'iterations': 914, 'depth': 4, 'learning_rate': 0.08970569382642916, 'l2
_leaf_reg': 5.208914529555241, 'random_strength': 0.6056908405757965, 'threshold':
0.5265637446946859}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:39:04,383] Trial 34 finished with value: 0.6894246657653599 and pa
rameters: {'iterations': 885, 'depth': 4, 'learning_rate': 0.08563542476945872, 'l2
_leaf_reg': 1.9698823210877507, 'random_strength': 0.6014485322373097, 'threshold':
0.4747036570744317}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:39:26,228] Trial 35 finished with value: 0.690051186598418 and par
ameters: {'iterations': 903, 'depth': 4, 'learning_rate': 0.08857510695584357, 'l2_
leaf_reg': 5.344369944434175, 'random_strength': 0.6127585773294119, 'threshold':
0.5216220125955369}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:39:47,960] Trial 36 finished with value: 0.6817891373801916 and pa
rameters: {'iterations': 879, 'depth': 4, 'learning_rate': 0.09212868852239235, 'l2
_leaf_reg': 5.160444121912242, 'random_strength': 0.6676284503271369, 'threshold':
0.5282485643250683}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:40:08,060] Trial 37 finished with value: 0.6510180918245434 and pa
rameters: {'iterations': 806, 'depth': 4, 'learning_rate': 0.1180814405698567, 'l2_
leaf_reg': 1.5327246913753898, 'random_strength': 0.6033498070027341, 'threshold':
0.6428788359871643}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:40:31,314] Trial 38 finished with value: 0.6735943223443224 and pa
rameters: {'iterations': 961, 'depth': 4, 'learning_rate': 0.05626332052947809, 'l2
_leaf_reg': 2.7707208241682624, 'random_strength': 0.6790668809396273, 'threshold':
0.5202764795395739}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:40:52,816] Trial 39 finished with value: 0.6616988416988415 and pa
rameters: {'iterations': 880, 'depth': 4, 'learning_rate': 0.13784115911312791, 'l2
_leaf_reg': 2.383750064658659, 'random_strength': 0.5763783977940353, 'threshold':
0.60218968826804}. Best is trial 32 with value: 0.6988019891500904.
[I 2025-12-19 23:41:13,591] Trial 40 finished with value: 0.7051092896174863 and pa
rameters: {'iterations': 858, 'depth': 4, 'learning_rate': 0.09041179707806742, 'l2
_leaf_reg': 4.197379121804284, 'random_strength': 0.715638441058104, 'threshold':
0.4871332648045998}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:41:34,898] Trial 41 finished with value: 0.6907634164777022 and pa
rameters: {'iterations': 868, 'depth': 4, 'learning_rate': 0.10910106761308455, 'l2
_leaf_reg': 4.243059967106836, 'random_strength': 0.7810960867308852, 'threshold':
0.47977559440767154}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:41:55,263] Trial 42 finished with value: 0.6772782738095238 and pa
rameters: {'iterations': 849, 'depth': 4, 'learning_rate': 0.1089419409638784, 'l2_
leaf_reg': 4.159218660747422, 'random_strength': 0.8347495463563874, 'threshold':
0.4844363784213526}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:42:18,357] Trial 43 finished with value: 0.7008562415502477 and pa
rameters: {'iterations': 771, 'depth': 5, 'learning_rate': 0.06358708091037772, 'l2
_leaf_reg': 1.5162220171147864, 'random_strength': 0.7308493019782363, 'threshold':
0.39514442349908036}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:42:41,953] Trial 44 finished with value: 0.6825061890199504 and pa
rameters: {'iterations': 775, 'depth': 5, 'learning_rate': 0.06081909855421148, 'l2
```

```
_leaf_reg': 1.5432831958649764, 'random_strength': 0.7225189470055566, 'threshold':
0.3998570048082379}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:43:12,730] Trial 45 finished with value: 0.6847920634920635 and pa
rameters: {'iterations': 965, 'depth': 5, 'learning_rate': 0.04341913793449644, 'l2
_leaf_reg': 4.26472559711239, 'random_strength': 0.7959958166088235, 'threshold':
0.5532646204866147}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:43:31,054] Trial 46 finished with value: 0.6847802197802197 and pa
rameters: {'iterations': 738, 'depth': 4, 'learning_rate': 0.1712462707363745, 'l2_
leaf_reg': 3.3566663925489713, 'random_strength': 0.7065749291856556, 'threshold':
0.397254417675055}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:43:56,660] Trial 47 finished with value: 0.6835913978494623 and pa
rameters: {'iterations': 814, 'depth': 5, 'learning_rate': 0.21600662356041758, 'l2
_leaf_reg': 5.669653659169295, 'random_strength': 0.917013481020252, 'threshold':
0.45197864843017377}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:44:17,997] Trial 48 finished with value: 0.685764910615409 and par
ameters: {'iterations': 853, 'depth': 4, 'learning_rate': 0.06241411463211009, 'l2_
leaf_reg': 4.841065462996363, 'random_strength': 0.9979565286664225, 'threshold':
0.5096640059060266}. Best is trial 40 with value: 0.7051092896174863.
[I 2025-12-19 23:44:42,887] Trial 49 finished with value: 0.6450757575757575 and pa
rameters: {'iterations': 1008, 'depth': 4, 'learning_rate': 0.07639866832668434, 'l
2_leaf_reg': 2.4528121946600585, 'random_strength': 0.6515273288679068, 'threshol
d': 0.7017761127637772}. Best is trial 40 with value: 0.7051092896174863.
```

In [41]:
```
study_cb.best_params
```

Out[41]:
```
{'iterations': 858,
 'depth': 4,
 'learning_rate': 0.09041179707806742,
 'l2_leaf_reg': 4.197379121804284,
 'random_strength': 0.715638441058104,
 'threshold': 0.4871332648045998}
```

Гиперпараметры были успешно подобраны

Обучение Catboost с подобранными гиперпармтерами

In [42]:
```
best_params = study_cb.best_params

cb_tuned = CatBoostClassifier(
    iterations=best_params["iterations"],
    depth=best_params["depth"],
    learning_rate=best_params["learning_rate"],
    l2_leaf_reg=best_params["l2_leaf_reg"],
    random_strength=best_params["random_strength"],
    loss_function="Logloss",
    eval_metric="AUC",
    verbose=0,
    random_state=42
)

cb_tuned.fit(
    X_train_cb,
    y_train,
    cat_features=cat_feature_indices,
    verbose=False
)
```

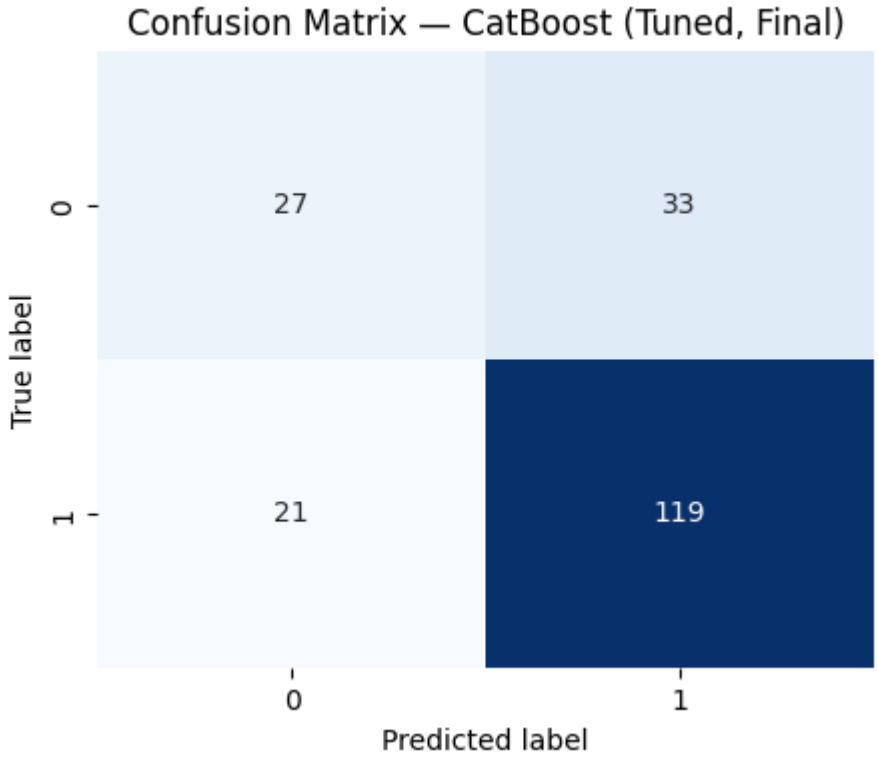
Out[42]: <catboost.core.CatBoostClassifier at 0x190835b0f80>

Оцениваем метрики качества

In [43]:
```
y_proba_cb_test = cb_tuned.predict_proba(X_test_cb)[:, 1]

evaluate_classification_model(
    y_true=y_test,
    y_proba=y_proba_cb_test,
    threshold=best_params["threshold"],
    model_name="CatBoost (Tuned, Final)"
)
```

```
CatBoost (Tuned, Final)
-----------------------
ROC-AUC : 0.7645
F1-score: 0.8151
Threshold: 0.4871332648045998
```



Out[43]:
```
{'model': 'CatBoost (Tuned, Final)',
 'roc_auc': 0.7645238095238096,
 'f1_score': 0.815068493150685,
 'threshold': 0.4871332648045998}
```

CatBoost показывает стабильные и конкурентоспособные результаты, подтверждая целесообразность использования градиентного бустинга для задачи кредитного скоринга.

Сравнение всех моделей между собой

In [50]:
```
models_metrics = [
    {
        "Model": "Logistic Regression (Tuned)",
```

```
        "ROC-AUC": 0.7592,
        "F1-score": 0.7817,
        "Threshold": 0.30
    },
    {
        "Model": "Random Forest (Tuned)",
        "ROC-AUC": 0.7881,
        "F1-score": 0.8333,
        "Threshold": 0.40
    },
    {
        "Model": "CatBoost (Tuned)",
        "ROC-AUC": 0.7645,
        "F1-score": 0.8151,
        "Threshold": 0.49
    }
]
metrics_df = pd.DataFrame(models_metrics)
metrics_df
```

Out[50]:

	Model	ROC-AUC	F1-score	Threshold
0	Logistic Regression (Tuned)	0.7592	0.7817	0.30
1	Random Forest (Tuned)	0.7881	0.8333	0.40
2	CatBoost (Tuned)	0.7645	0.8151	0.49

Random Forest является победителем. Он выдает самые лучшие качества на тестовых выборках.

Сохранение моделей

In [46]:

```
rf_best_params = study_rf.best_params

rf_artifact = {
    "model": rf_tuned,
    "preprocessor": rf_preprocessor,
    "best_params": {
        "n_estimators": rf_best_params["n_estimators"],
        "max_depth": rf_best_params["max_depth"],
        "min_samples_split": rf_best_params["min_samples_split"],
        "min_samples_leaf": rf_best_params["min_samples_leaf"],
        "max_features": rf_best_params["max_features"],
        "threshold": rf_best_params["threshold"]
    },
    "metrics": {
        "roc_auc": 0.7881,
        "f1_score": 0.8333
    }
}

joblib.dump(
    rf_artifact,
    "models/rf_tuned_model.joblib"
)
```

Out[46]: ['models/rf_tuned_model.joblib']

In [47]:
```
cb_artifact = {
    "model": cb_tuned,
    "cat_feature_indices": cat_feature_indices,
    "best_params": {
        "iterations": study_cb.best_params["iterations"],
        "depth": study_cb.best_params["depth"],
        "learning_rate": study_cb.best_params["learning_rate"],
        "l2_leaf_reg": study_cb.best_params["l2_leaf_reg"],
        "random_strength": study_cb.best_params["random_strength"],
        "threshold": study_cb.best_params["threshold"]
    },
    "metrics": {
        "roc_auc": 0.7645,
        "f1_score": 0.8151
    }
}

joblib.dump(
    cb_artifact,
    "models/catboost_tuned_model.joblib"
)
```

Out[47]: ['models/catboost_tuned_model.joblib']

Основные выводы

- Совместная оптимизация гиперпараметров и порога классификации существенно улучшает качество моделей на уровне бизнес-решений.
- Использование комбинированной метрики F1 × ROC-AUC позволяет избежать перекоса в сторону одной из характеристик качества.
- **Random Forest** был выбран в качестве финальной модели кредитного скоринга благодаря наилучшему общему качеству на тестовой выборке.
- Все обученные модели были сохранены вместе с параметрами и порогами, что обеспечивает воспроизводимость и готовность к дальнейшему использованию.