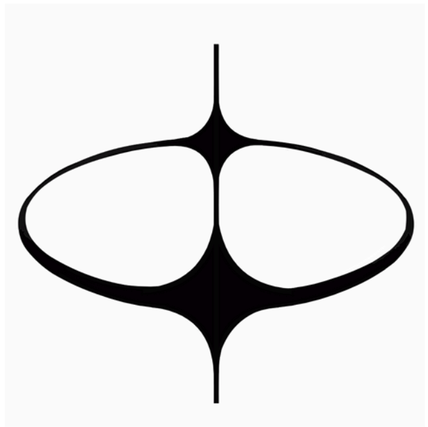

O'ZBEKISTON RESPUBLIKASI
MARKAZIY BANKI



Сводный анализ

1 Объединение и очистка данных

- Интеграция 6 источников → 90,000 строк
- Нормализация форматов, удаление шумовых признаков

2 Feature Engineering 1

- Создание 25 производных признаков
- Out-of-Fold методология с 5-fold кросс-валидацией

3 Feature Engineering 2

- KNN meta-features
- WOE-трансформация для монотонности зависимостей
- Polynomial & Interaction features для нелинейных паттернов

4 Обучение и результаты

- scale_pos_weight решение дисбаланса классов
- Bayesian метод - 30 trials.
- Финальная модель - это ensemble из трёх XGBoost с разными параметрами и оптимизированными весами.
- Добавили Isotonic калибровку для корректных вероятностей

AUC 0.8

Ключевые операции очистки:

Форматирование
- Удаление "$" и "," из финансовых полей → конвертация в float
- Нормализация категорий (Full-time, FULL_TIME → Full-Time)
- Преобразование дат в числовые признаки

Обработка пропусков
- JSONL: 89,999 записей → заполнение медианами
- XML: 89,999 записей → интерполяция

Удаление шума
- random_noise_1 - корреляция ~0
- referral_code - 80K+ уникальных значений
- previous_zip_code - низкая предсказательная сила
- Дубликаты: application_id, customer_ref после merge

Feature Engineering 1: Финансовые метрики - 24 признака

1 фин метрики

Колонка	Формула	Описание
calculated_dti	total_debt_amount / annual_income	Пересчитанный DTI
calculated_pti	(monthly_payment + existing_monthly_debt) / monthly_income	Пересчитанный PTI
calculated_utilization	credit_usage_amount / total_credit_limit	Пересчитанная утилизация
debt_payment_burden	(monthly_payment + existing_monthly_debt) / monthly_income	Бремя долга
total_debt_to_income_annual	total_debt_amount / annual_income	Годовой DTI
total_monthly_debt_payment	monthly_payment + existing_monthly_debt	Общий месячный платеж
annual_debt_payment	total_monthly_debt_payment * 12	Годовой платеж по долгу
interest_burden	(interest_rate / 100) * loan_amount / monthly_income	Бремя процентов

2 ликвидность

Колонка	Формула	Описание
disposable_income	monthly_income - total_monthly_debt_payment	Располагаемый доход
monthly_free_cash_flow	monthly_income - total_monthly_debt_payment - (loan_amount/loan_term)	Свободный денежный поток
free_cash_flow_ratio	monthly_free_cash_flow / monthly_income	Коэффициент свободного потока
income_to_payment_capacity	monthly_income / (monthly_payment + 0.01)	Способность платить

4 доходы

Колонка	Формула	Описание
monthly_income_from_annual	annual_income / 12	Месячный доход из годового
income_to_regional_median	annual_income / regional_median_income	Доход к региональной медиане
income_source_match	1 if abs(monthly_income - monthly_income_from_annual) < 500 else 0	Совпадение источников дохода

3 кредитное поведение

Колонка	Формула/Источник	Описание
account_diversity_index	num_credit_accounts / (credit_history_depth + 1)	Индекс разнообразия счетов
delinquency_rate	num_delinquencies_2yrs / (num_credit_accounts + 1)	Уровень просрочек
num_delinquencies_2yrs	Из credit_history	Просрочки за 2 года
num_public_records	Из credit_history	Публичные записи
negative_marks_total	num_delinquencies_2yrs + num_public_records	Всего негативных отметок
oldest_account_age_months	account_open_year преобразован	Возраст старейшего счета
service_call_intensity	num_customer_service_calls / (oldest_account_age_months + 1)	Интенсивность звонков

5 региональность

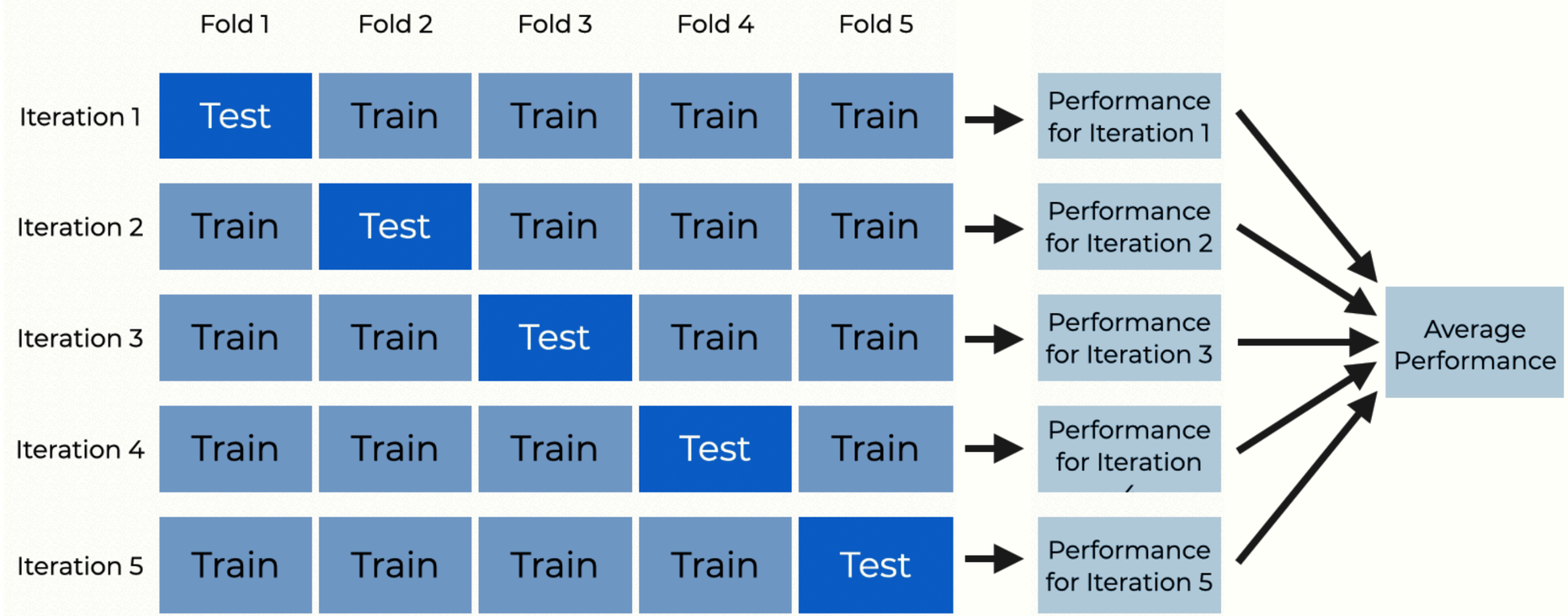
Колонка	Формула	Описание
housing_affordability	(regional_median_rent * 12) / annual_income	Доступность жилья
regional_stress_index	regional_unemployment_rate * cost_of_living_index	Региональный стресс

Признаки feature eng 2

№	Колонка	Категория	Формула/Метод
1	knn_target_prob_50	KNN Meta	KNN(k=50).predict_proba()
2	knn_target_prob_100	KNN Meta	KNN(k=100).predict_proba()
3	knn_target_prob_500	KNN Meta	KNN(k=500).predict_proba()
4	state_target_encoded	Target Encoding	Out-of-fold mean(default) по state
5	debt_credit_interaction	Interaction	debt_to_income_ratio *
6	age_credit_interaction	Interaction	age * credit_score / 100
7	credit_stress_score_squared	Polynomial	credit_stress_score²
8	credit_stress_score_cbrt	Polynomial	∛credit_stress_score
9	debt_to_income_ratio_squared	Polynomial	debt_to_income_ratio²
10	debt_to_income_ratio_cbrt	Polynomial	∛debt_to_income_ratio
11	credit_utilization_squared	Polynomial	credit_utilization²
12	credit_utilization_cbrt	Polynomial	∛credit_utilization
13	employment_income_stability	Composite	employment_length * monthly_income

Feature Engineering v2: 13 признаков

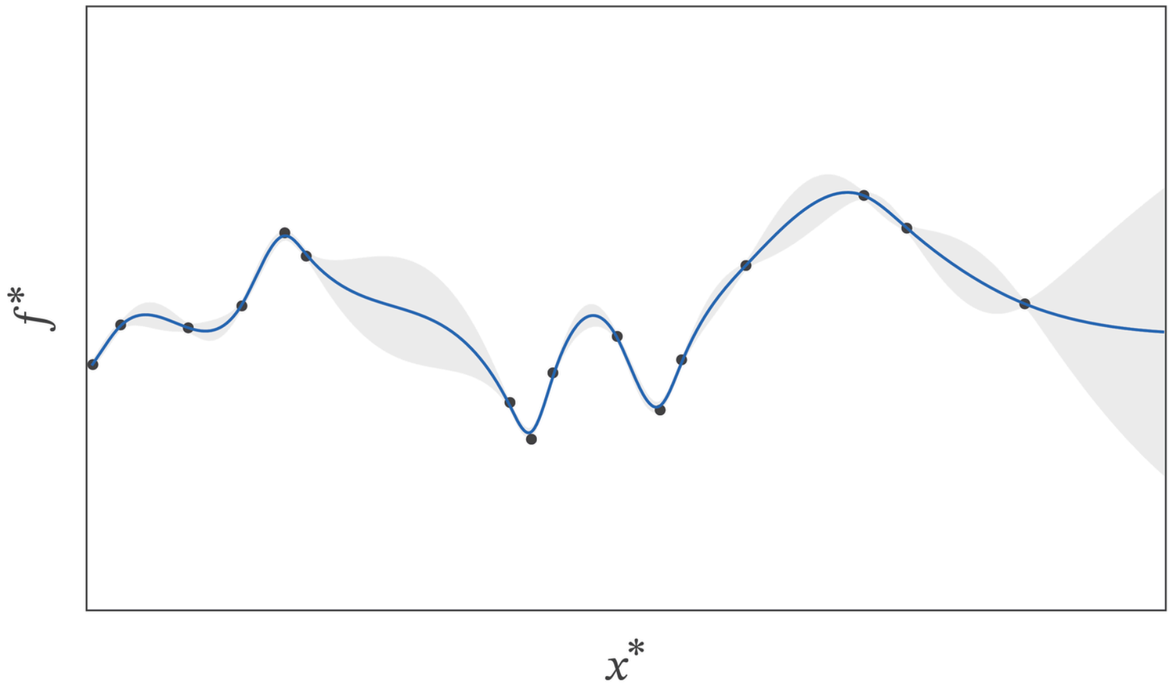
Признаки feature eng 2



№			
11	credit_utilization_squared	Polynomial	$credit\_utilization^2$
12	credit_utilization_cbrt	Polynomial	$\sqrt[3]{credit\_utilization}$
13	employment_income_stability	Composite	employment_length * monthly_income

Байесовская оптимизация

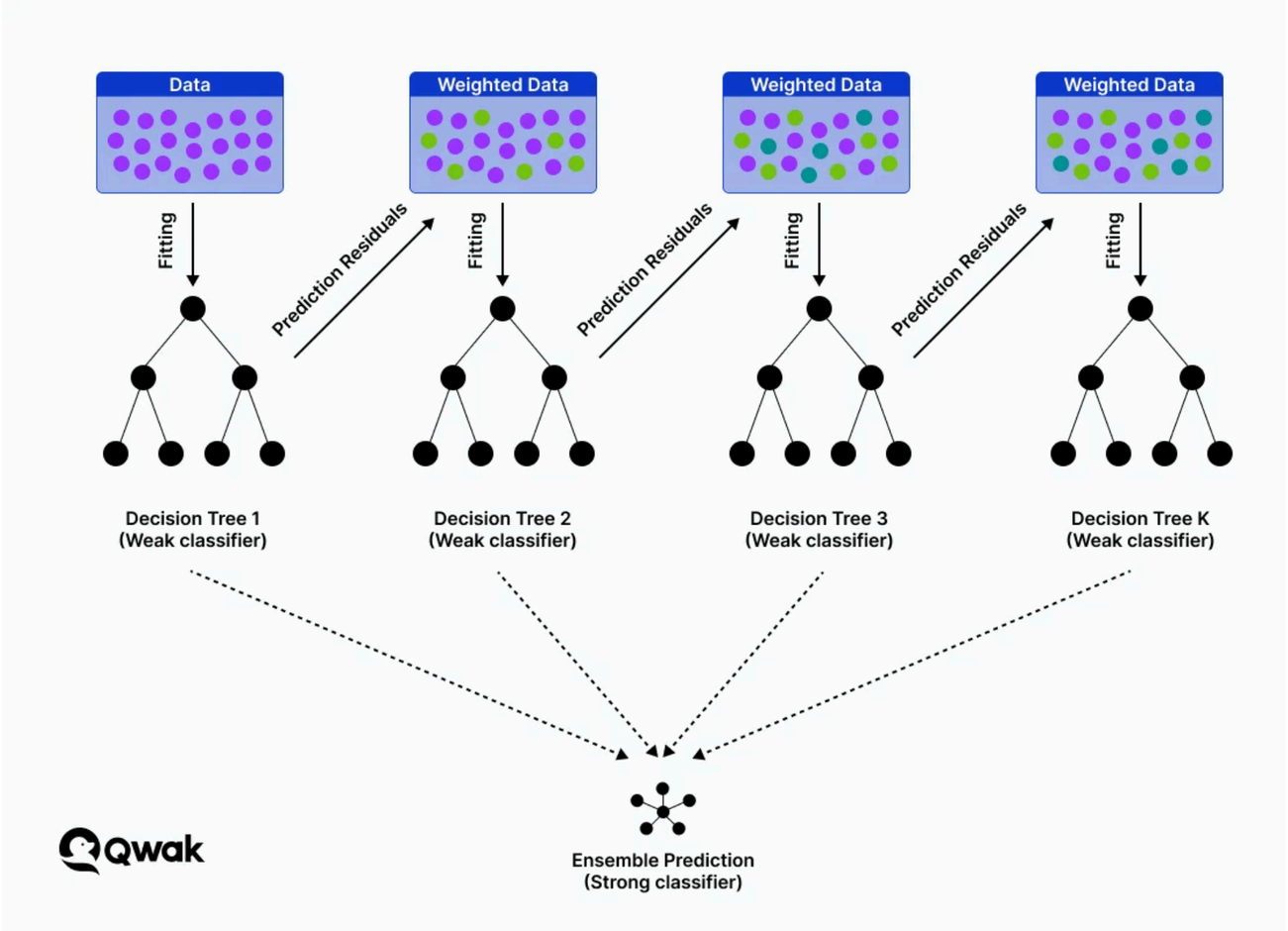
это подход к настройке гиперпараметров в моделях машинного обучения, особенно эффективный для ресурсоёмких целевых функций. В отличие от более простых методов, таких как поиск по сетке или случайный поиск, он использует вероятностную модель для управления поиском оптимальных гиперпараметров.



Обучение: Ensembling xgboost

Ensembling method

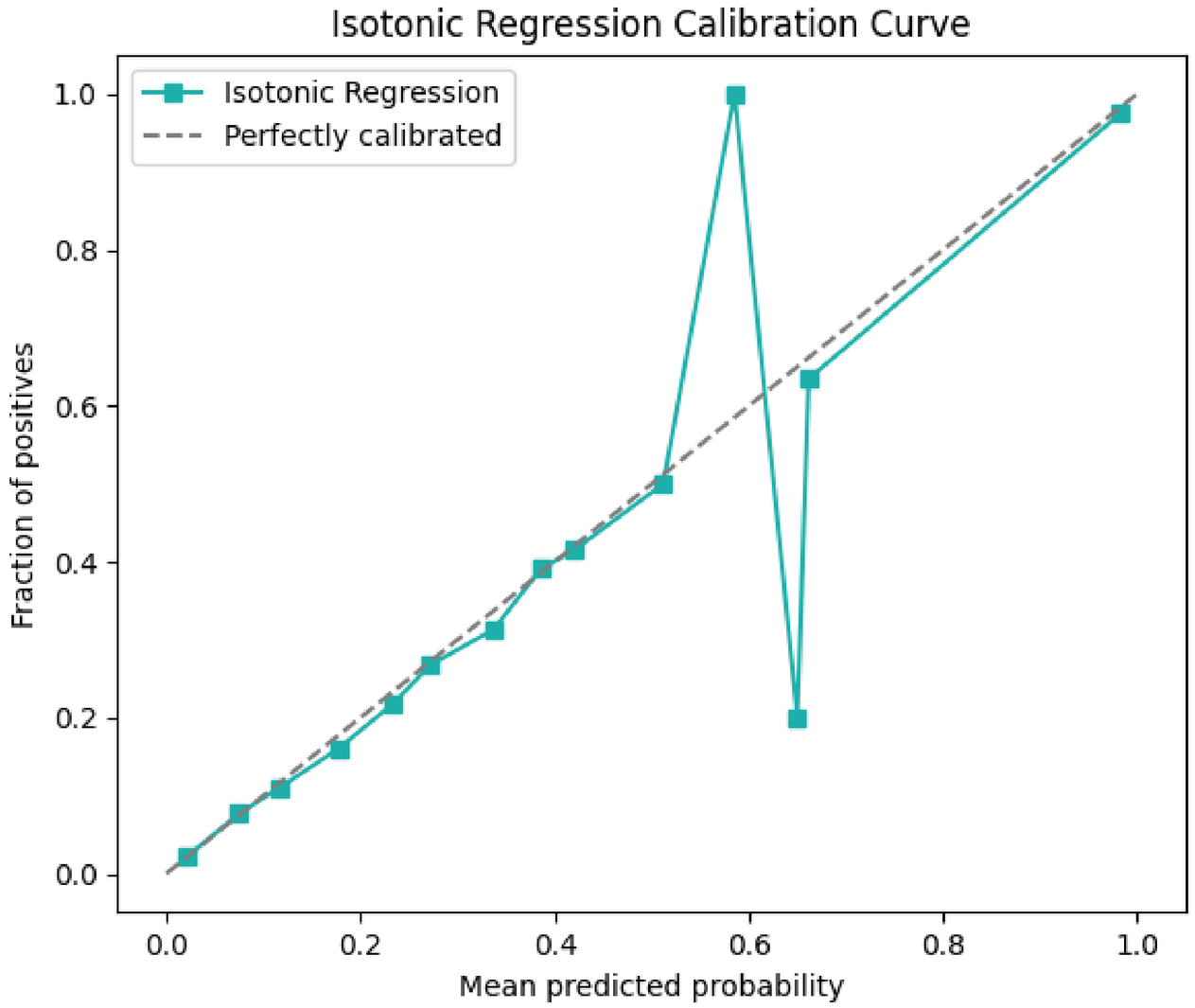
1. Начальный прогноз:
2. Вычисление остатков (ошибок)
3. Последовательное обучение деревьев: Новое дерево решений обучается на этих остатках (ошибках) с использованием метода оптимизации
4. Обновление прогноза: Прогноз нового дерева добавляется к общему прогнозу ансамбля. Каждое новое дерево вносит небольшой вклад, что контролируется гиперпараметром
5. Повторение: Процесс повторяется, пока не будет достигнут желаемый уровень производительности или максимальное количество деревьев не будет построено.



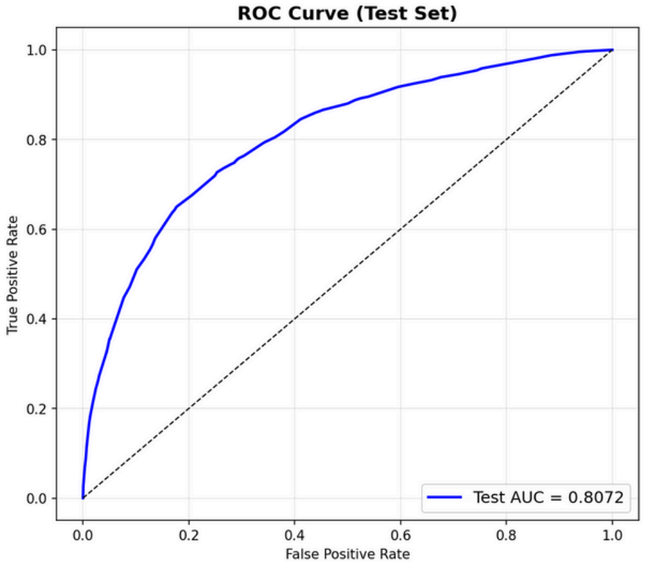
Обучение: isotonic calibration

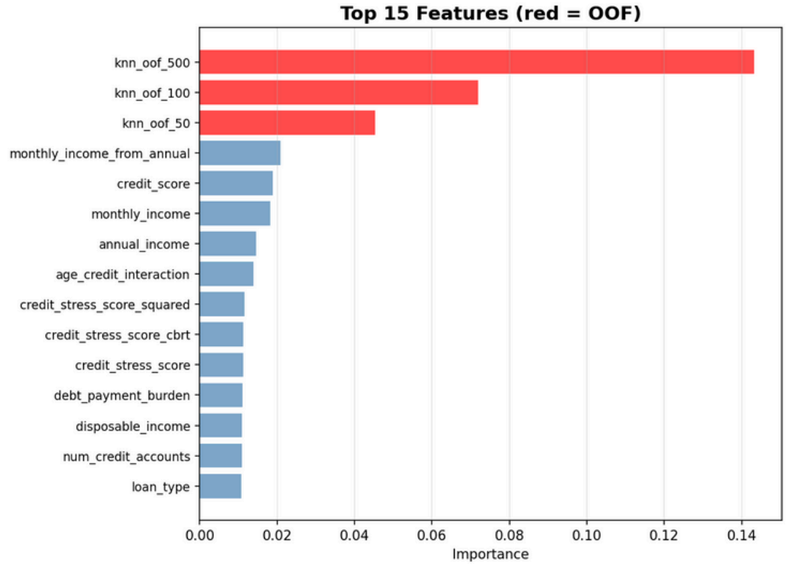
Isotonic Calibration

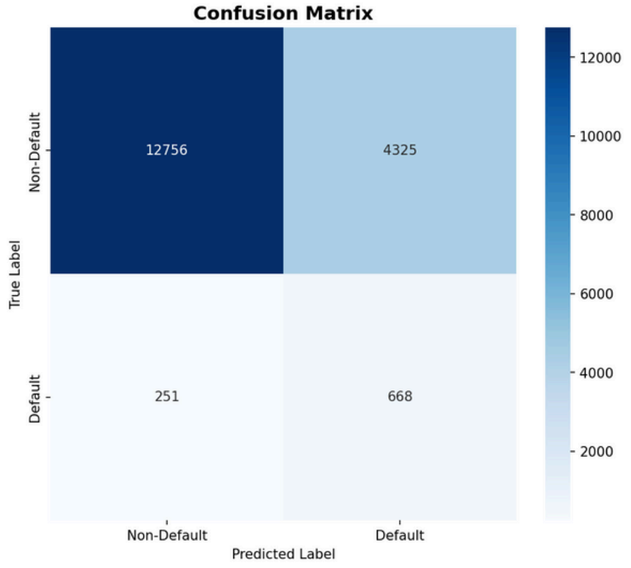
isotonic regression or monotonic regression is the technique of fitting a free-form line to a sequence of observations

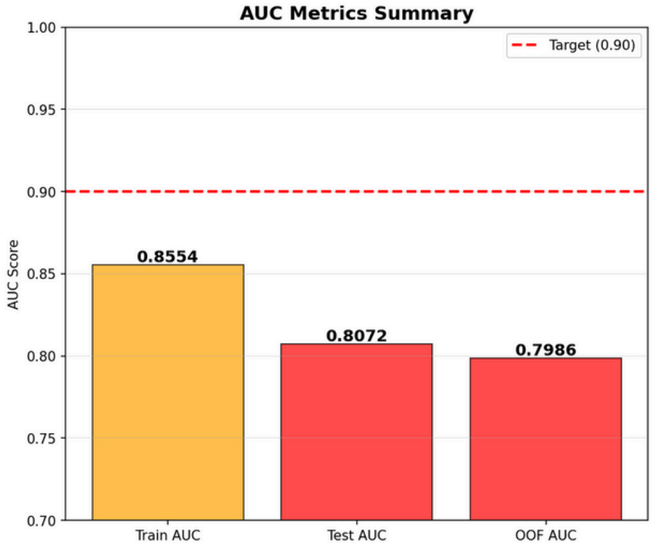


Результаты


ROC Curve (Test Set)
True Positive Rate
False Positive Rate
Test AUC = 0.8072


Top 15 Features (red = OOF)
knn_oof_500
knn_oof_100
knn_oof_50
monthly_income_from_annual
credit_score
monthly_income
annual_income
age_credit_interaction
credit_stress_score_squared
credit_stress_score_cbrt
credit_stress_score
debt_payment_burden
disposable_income
num_credit_accounts
loan_type
Importance


Confusion Matrix
12756
4325
251
668
Non-Default
Default
True Label
Predicted Label


AUC Metrics Summary
Target (0.90)
AUC Score
0.8554
0.8072
0.7986
Train AUC
Test AUC
OOF AUC