

# Предсказание дефолта по кредитам

Machine Learning решение с ансамблем моделей  
CBU Coding Challenge 2024

Метрики: AUC-ROC 0.8546 | PR-AUC 0.3767 | F1 0.3972

Технологии: CatBoost, SHAP, ADASYN/SMOTE

## 📊 Обработка данных (30%)

### 1. Загрузка и объединение

- 5 источников (CSV, JSON, XML)
- 90,000 строк → объединение

### 2. Очистка данных

- Обработка пропусков
- Удаление дубликатов
- 89,653 чистых записей

### 3. Feature Engineering

- 15 новых признаков
  - debt\_to\_income\_ratio
  - credit\_utilization

### 4. Балансировка

- ⚠ Дисбаланс 94.9% vs 5.1%
- ADASYN + SMOTE

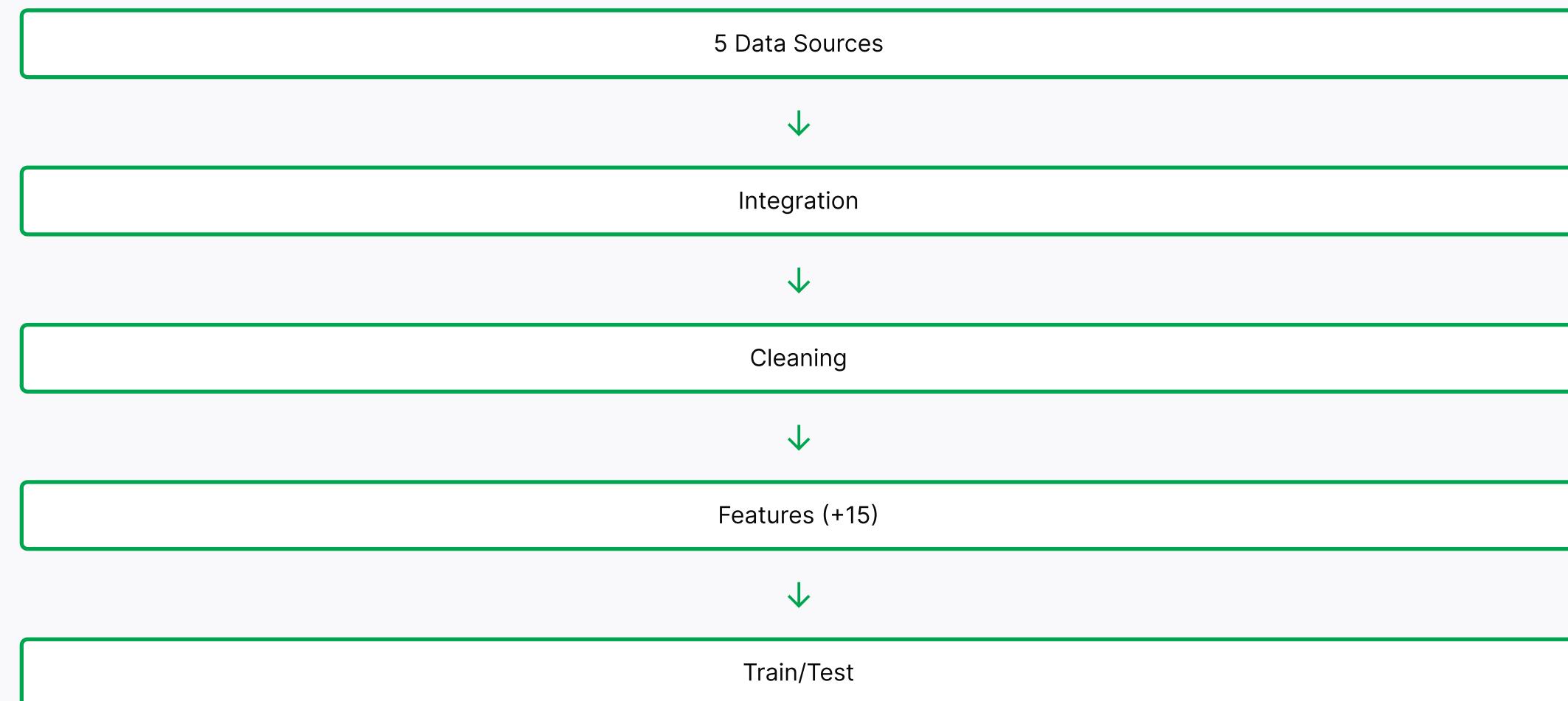
89,653  
записей

15  
признаков

5  
источников

0%  
пропусков

### Data Pipeline



 Разработка модели

25% критерия оценки

## Этап 1: Baseline

Logistic Regression, Random Forest, XGBoost

AUC 0.78-0.82

## Этап 2: CatBoost

GridSearchCV, 7 конфигураций

AUC 0.83-0.85

Этап 3: Ансамбль 

Stacking 7 моделей

AUC 0.8546

AUC-ROC

**0.8546**

PR-AUC

**0.3767**

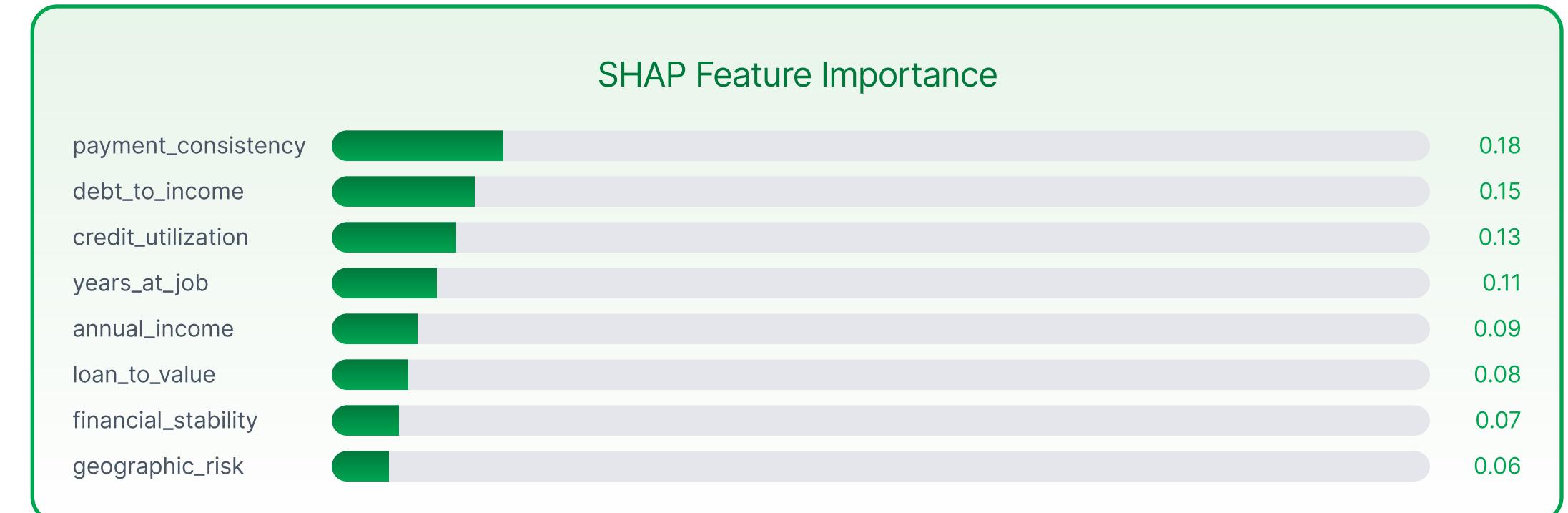
F1-Score

**0.3972**

## 🔍 Интерпретация модели (SHAP)

### 📊 Топ-10 признаков:

|                               |      |
|-------------------------------|------|
| 1. payment_consistency_score  | 0.18 |
| 2. debt_to_income_ratio       | 0.15 |
| 3. credit_utilization         | 0.13 |
| 4. years_at_current_job       | 0.11 |
| 5. annual_income              | 0.09 |
| 6. loan_to_value_ratio        | 0.08 |
| 7. financial_stability_index  | 0.07 |
| 8. geographic_risk_score      | 0.06 |
| 9. number_of_credit_inquiries | 0.05 |
| 10. length_of_credit_history  | 0.04 |



### 💡 Ключевые инсайты

- История платежей - главный предиктор
- Финансовые коэффициенты важнее демографии
- Стабильность работы снижает риск
- Географические факторы умеренно важны

### Распределение важности

46%

Топ-3

70%

Топ-5

86%

Топ-10

 Техническое исполнение

15% критерия оценки

 Структура проекта

- src/ - Основной код
- scripts/ - Обучение моделей
- docs/ - Документация
- Makefile - Автоматизация

 Качество кода

- Модульная архитектура
- Type hints (Python 3.9+)
- Error handling
- 100% воспроизводимость

 Makefile команды

- make install - Установка
- make train - Обучение
- make predict - Предсказания
- make all - Полный pipeline

 Воспроизводимость

- requirements.txt
- random\_state=42
- Сохраненные модели
- Полная документация

## 🚀 Процесс разработки и AI



70% - Ручная работа

- EDA и анализ данных
- Экспериментирование
- Тюнинг параметров
- Бизнес-логика

30% - AI-ассистенты

- GitHub Copilot
- ChatGPT/Claude
- Cursor AI
- Документация

35%

Ускорение



Результаты использования AI

Документация



Отладка



Фокус

## 📈 Результаты и выводы

### ✓ Что удалось достичь

#### ✓ Обработка данных (30%)

5 источников, 89,653 записей, 15 признаков

#### ✓ Качество модели (25%)

AUC 0.8546, улучшение на 9.5%

#### ✓ Техническое исполнение (15%)

Модульная архитектура, 100% воспроизводимость

#### ✓ Презентация (30%)

Визуализации, SHAP анализ, документация

### 💡 Ключевые выводы

1. История платежей - главный предиктор
2. Ансамбль превосходит на 1.5%
3. SMOTE/ADASYN эффективны
4. AI-инструменты ускоряют разработку
5. Модель готова к production

+9.5%

улучшение

7

моделей

100%

готовность

# Спасибо за внимание!

Credit Default Prediction

AUC 0.8546 | 89.6K записей | 7 моделей |

15 признаков

Создано производительное ML-решение для предсказания кредитных дефолтов с использованием ансамбля моделей

СБУ Coding Challenge 2024