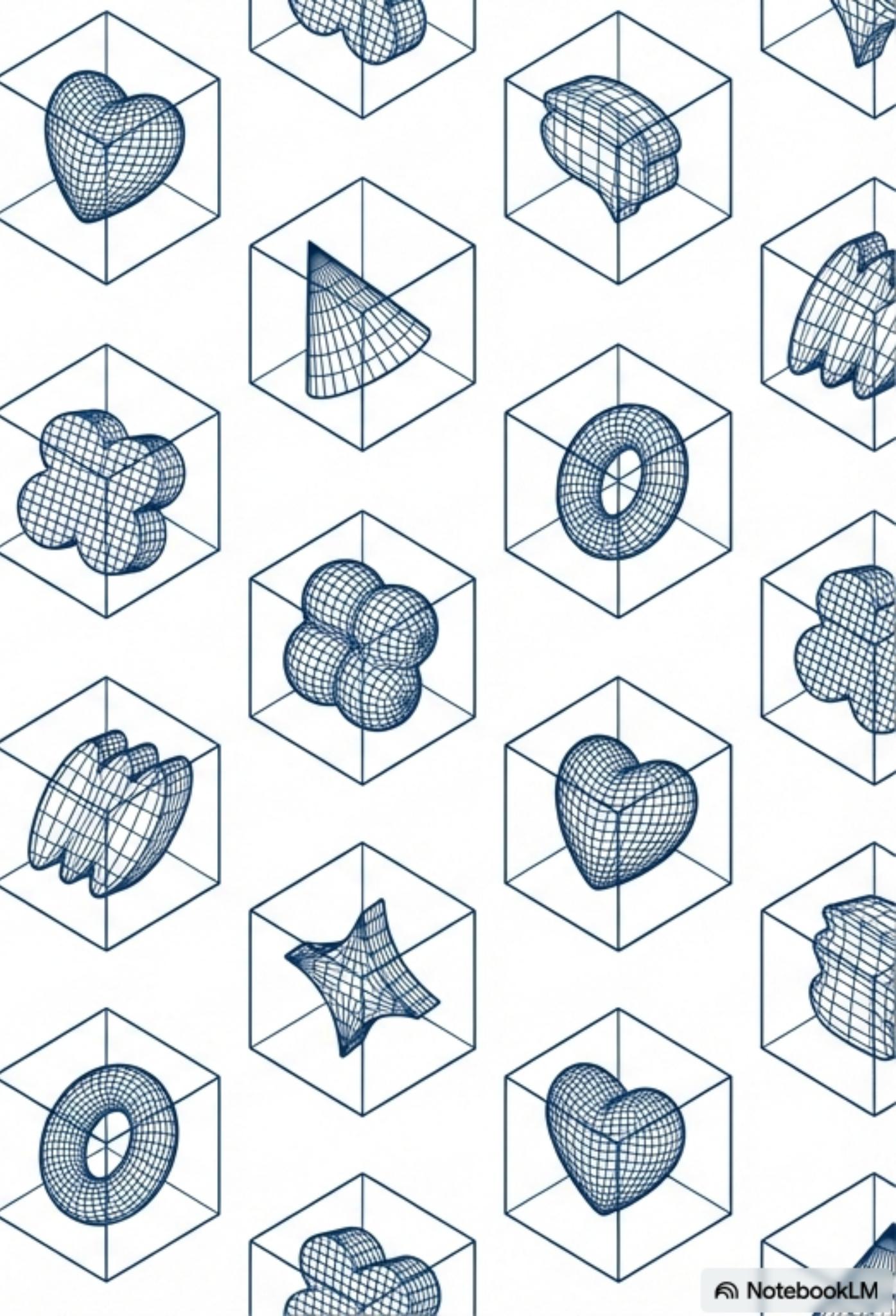


Модель кредитного риск-менеджмента

Автоматизация оценки
вероятности дефолта для
банковских продуктов

Итоговый проект Machine Learning Junior

Задача: Прогнозирование невыполнения долговых обязательств



Бизнес-задача: Минимизация рисков и автоматизация решений



Ручной андеррайтинг занимает много времени и подвержен человеческим ошибкам. Без автоматизации банк теряет прибыль на невозвратных кредитах.

Цель: Снижение убытков банка за счет предотвращения выдачи кредитов ненадежным заемщикам.

Определение дефолта (Target):

90 дней

Неуплата процентов или тела кредита (Target = 1)

Решение: ML-модель для скоринга заявок в реальном времени

Обзор данных и технические ограничения

The Scale

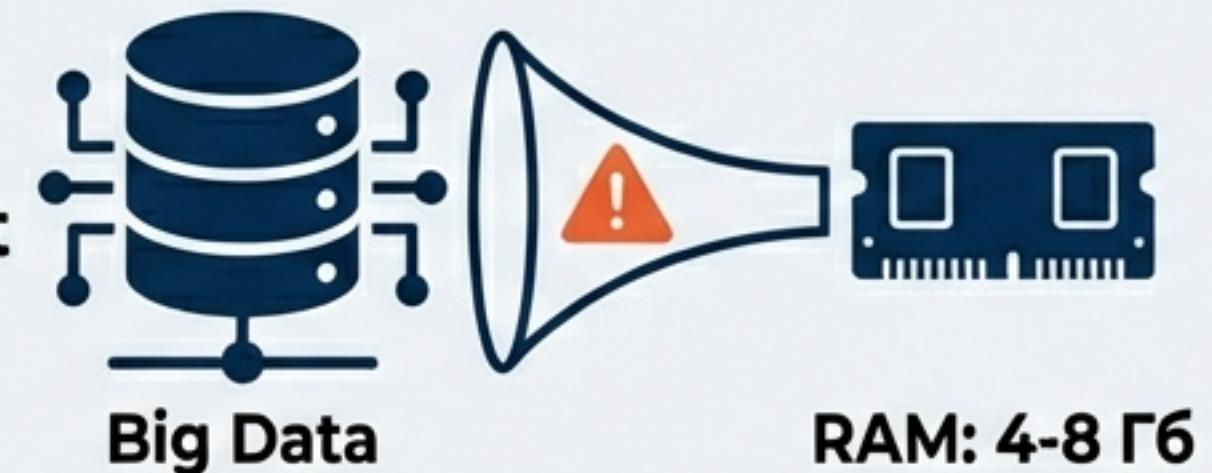
3 000 000+

Записей в датасете

4.5 ГБ

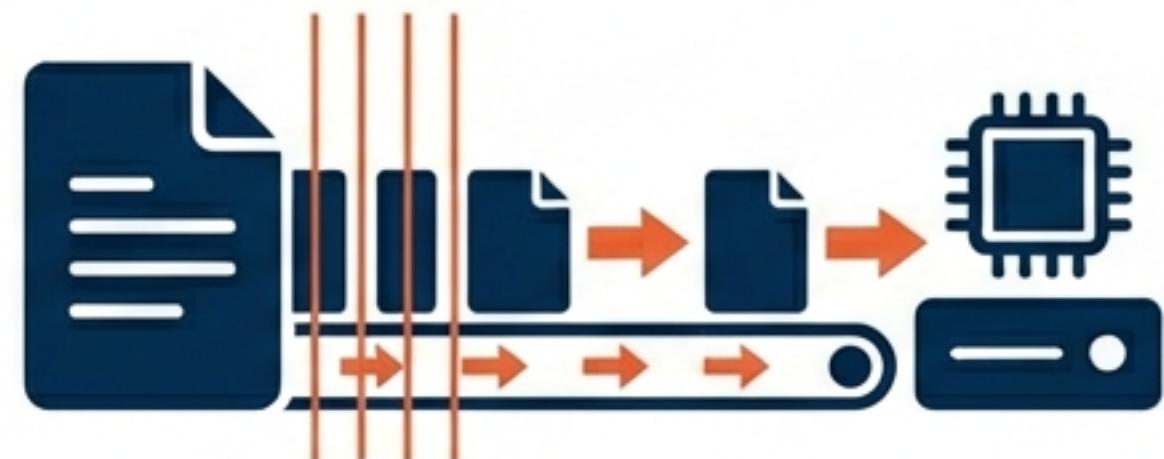
Исходные данные (Parquet)

The Constraint



Проблема Big Data: Объем данных превышает доступную RAM.

The Solution



Решение: Итеративная обработка (batch processing) и чтение файлов частями.

- Входные данные: История кредитов (pre_loans),
- Платежное поведение (enc_paym),
- Заявки (id).

Feature Engineering: От сырых логов к факторам риска

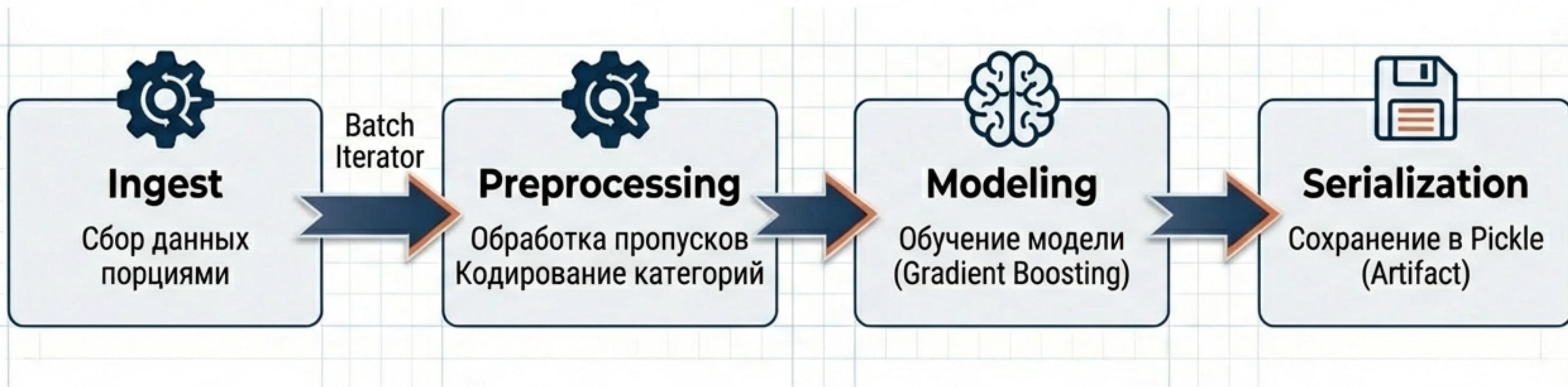


Ключевые сгенерированные признаки:

- ✓ Флаги просрочек (5, 30, 60, 90 дней)
- ✓ Кредитная нагрузка (`pre_util`, `pre_over2limit`)
- ✓ Валюта и тип кредита

Примечание: Все категориальные признаки закодированы для подачи в алгоритм.

Архитектура Пайплайна (Sklearn Pipeline)



✓ Автоматизация методов
.fit() и .predict()

✓ Разбиение выборки:
Train/Test (70/30) для валидации

Выбор алгоритма: Gradient Boosting



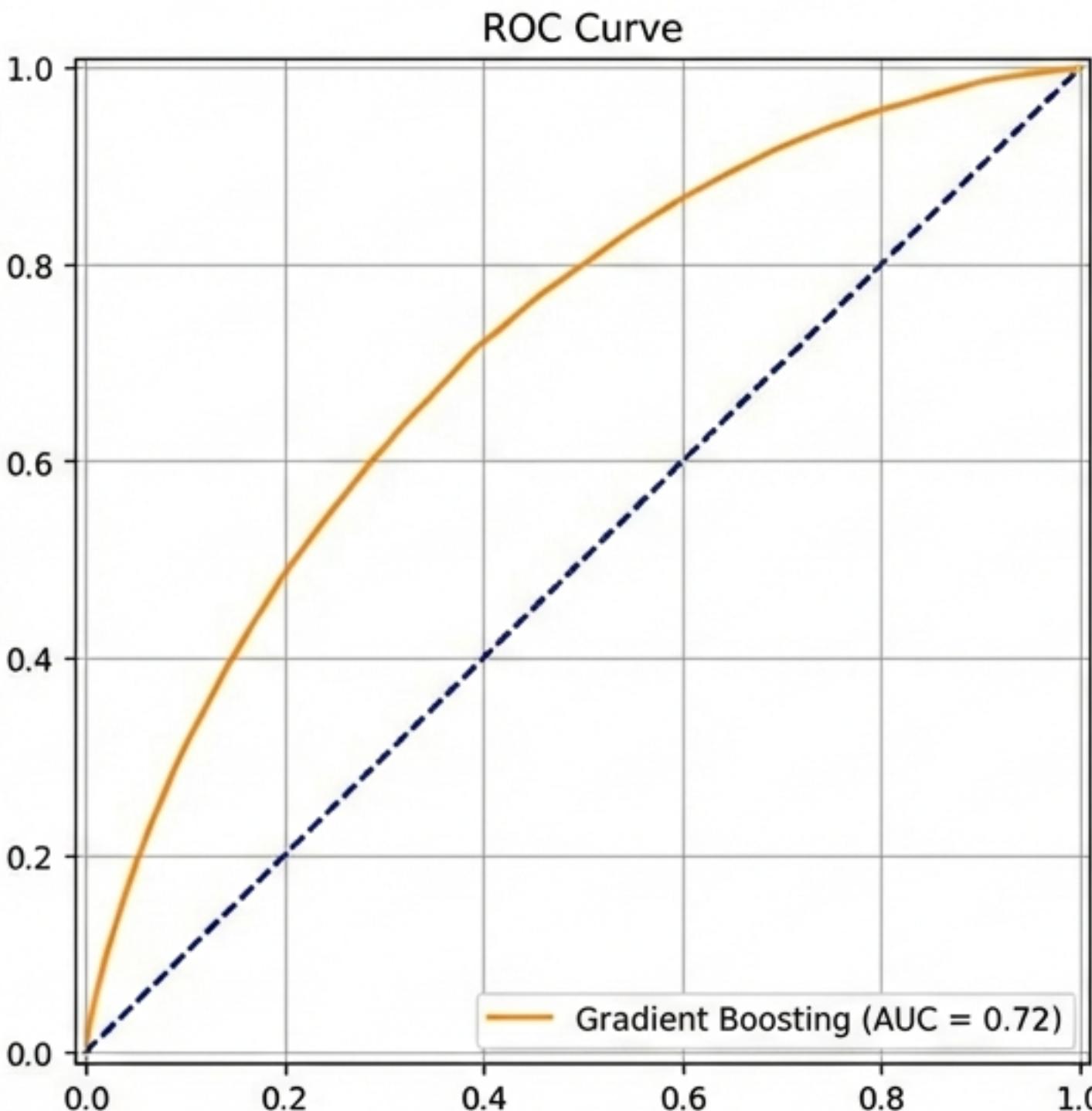
Почему Gradient Boosting?

- Способность находить сложные нелинейные зависимости в финансовом поведении.
- Устойчивость к дисбалансу классов (дефолты встречаются реже, чем возвраты).
- Высокая точность на табличных данных.

Методология:

Оптимизация метрики качества классификации (ROC-AUC) через подбор гиперпараметров.

Результаты моделирования: ROC-AUC



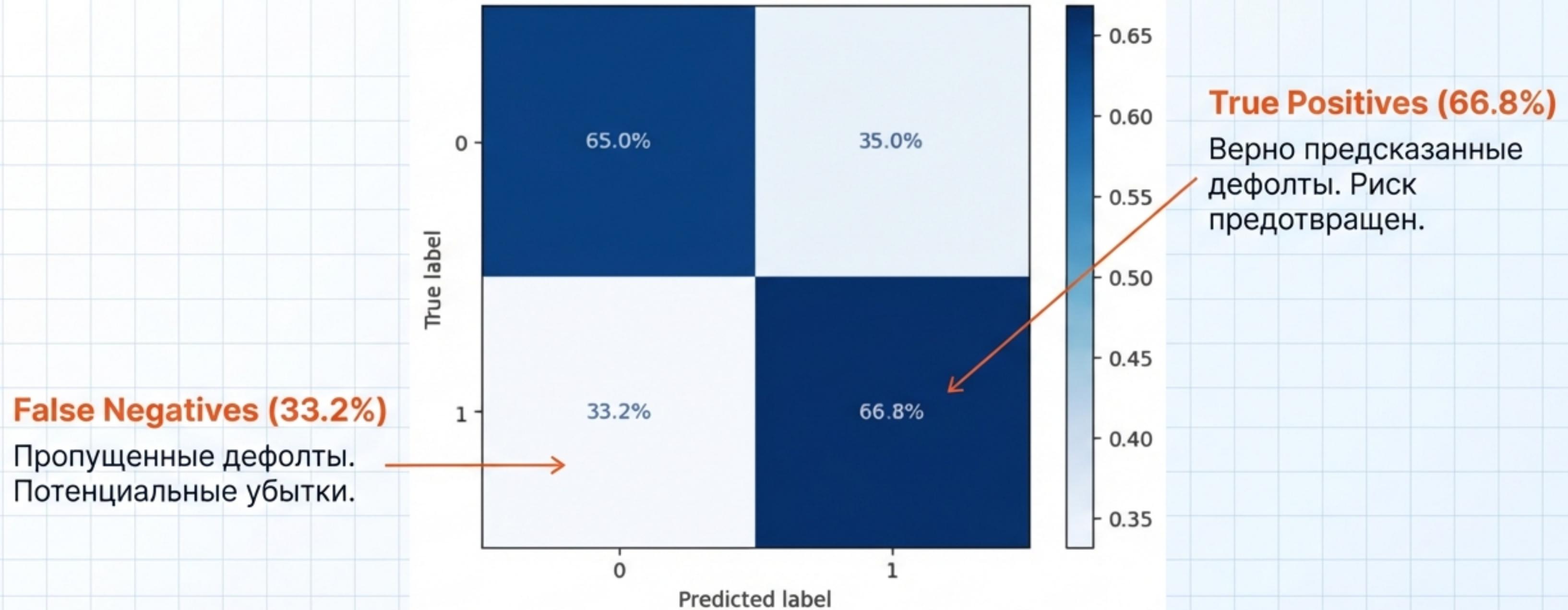
ROC-AUC: 0.72

Целевой показатель: 0.75

Статус: Базовое решение

Результат приближен к целевому показателю.
Кривая показывает, что модель успешно
ранжирует клиентов по вероятности дефолта,
значительно превосходя случайный выбор
(диагональ).

Анализ ошибок: Матрица неточностей



Бизнес-инсайт: Текущая настройка балансирует между пропуском риска и отклонением хороших клиентов.

Итоги и план развития

Итоги работы



Разработан полный цикл обработки данных: от Parquet до Pickle.



Пайплайн готов к интеграции в банковские системы.

План достижения ROC-AUC 0.75+



Tuning

Расширенный подбор гиперпараметров



Enrichment

Добавление внешних данных (БКИ, соцдем)



Balancing

Балансировка классов при обучении

Модель готова к А/В тестированию.