



Кредитный скоринг на данных German Credit


Анализ клиентов и модель прогноза риска дефолта


Мartiнович Георгий
<https://github.com/Ejtolf>

Проблема и цель

 **Проблема:** банки и лизинговые компании несут убытки из-за невозвратов кредитов.

 Решение часто принимается на основе поверхностных факторов => высокая вероятность ошибок.


 **Цель проекта:** построить модель машинного обучения, которая автоматизированно оценивает риск клиента (good / bad).


 Ожидаемый результат: снижение убытков и ускорение процесса одобрения заявок.





Гистограмма частот: размер кредита на good/bad клиентов

Данные

 **Размер выборки:** 1000 клиентов

 **Характеристики клиента:** возраст, пол, профессия (уровень квалификации), статус проживания, уровень сберегательного счёта, уровень расчётного счёта.

 **Характеристики кредита:** сумма кредита, срок, цель займа


 **Целевая переменная:** переменная риска (good / bad).


- У **18.4%** клиентов неизвестно состояние сберегательного счёта, у **39.4%** - состояние расчётного счёта.


	Unnamed: 0	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration		Purpose	Risk
0	0	67	male	2	own	NaN	little	1169	6		radio/TV	good
1	1	22	female	2	own	little	moderate	5951	48		radio/TV	bad
2	2	49	male	1	own	little	NaN	2096	12		education	good
3	3	45	male	2	free	little	little	7882	42	furniture/equipment		good
4	4	53	male	2	free	little	little	4870	24		car	bad


Фрагмент данных


Подход

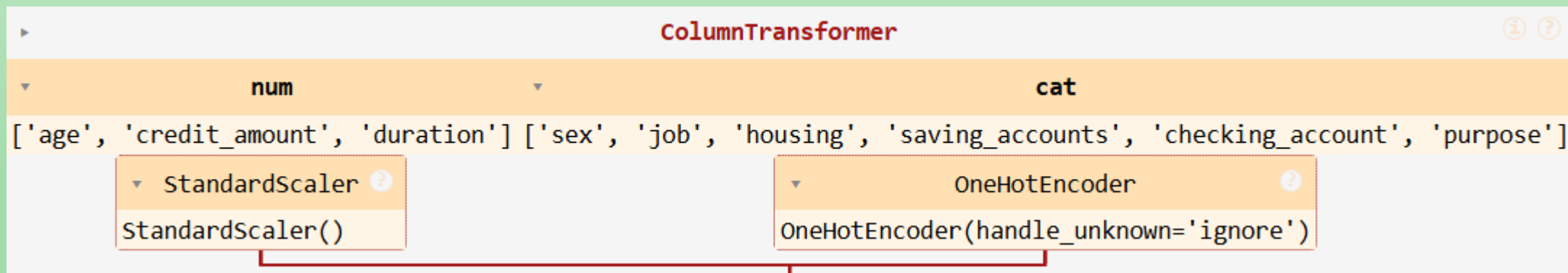
 **Предобработка данных:** очистка, обработка пропусков, кодирование категориальных признаков, добавлен новый признак (*ежемесячный платёж*).

 **Feature engineering:** масштабирование числовых признаков, one-hot encoding категориальных.

 **Моделирование:** использованы и протестированы три модели: **логистическая регрессия** (базовая интерпретируемая модель), **Random Forest** (хорошо работает с нелинейными зависимостями) и **Gradient Boost** (быстрая и точная на малом кол-ве данных)

 **Валидация:** 5-кратная стратифицированная кросс-валидация.

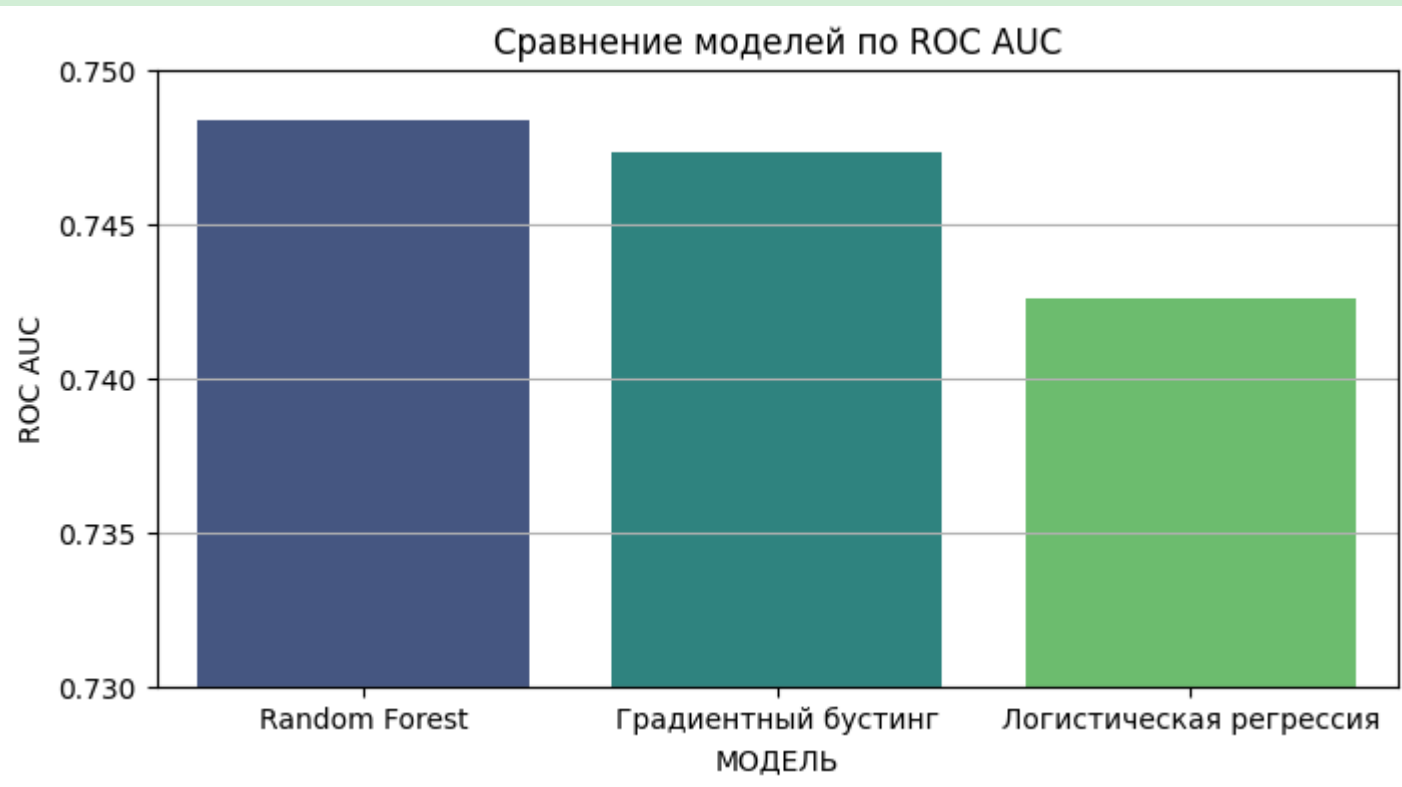
 **Метрика качества:** ROC-AUC (основная), а также Precision, Recall, F1.



Результаты

🎯 Метрика качества: ROC-AUC (насколько хорошо модель различает «надежных» и «рискованных» клиентов)

✅ Статистически НЕ важные признаки: работа (уровень квалификации) и цель кредита клиента.

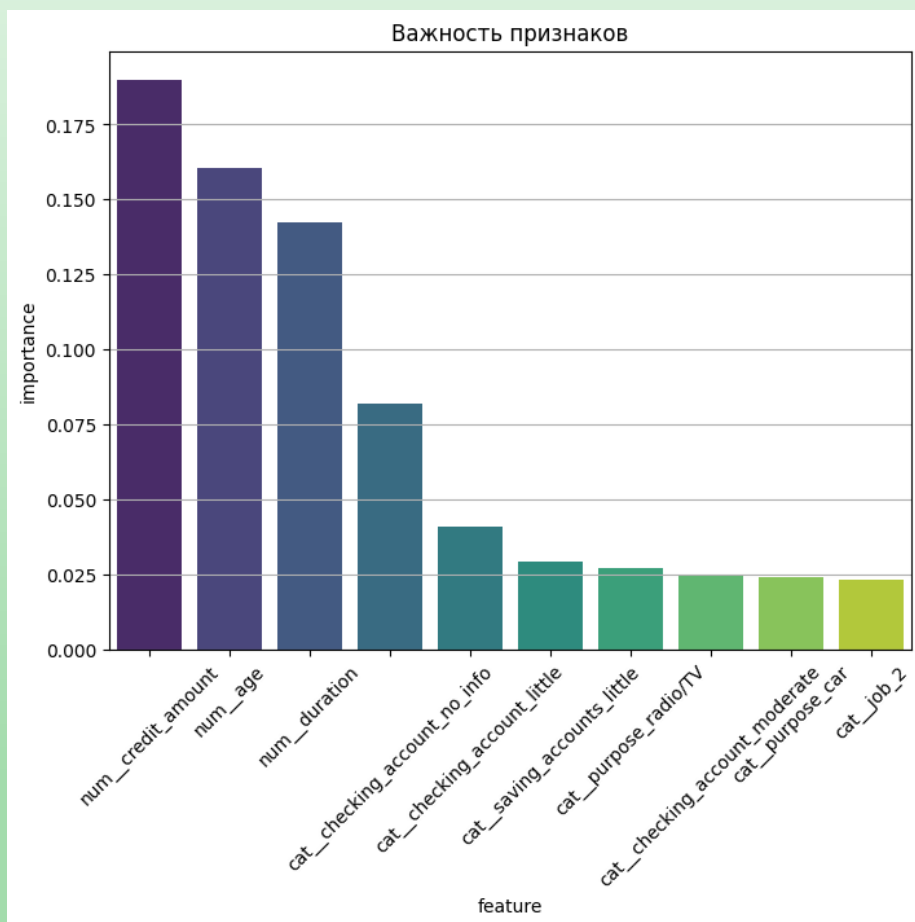


Лучшие результаты – Random Forest: **74.83%** правильных разделений на good/bad

Результаты

🎯 Метрика качества: ROC-AUC (насколько хорошо модель различает «надежных» и «рискованных» клиентов)

✅ Статистически НЕ важные признаки: работа (уровень квалификации) и цель кредита клиента.



💳 Крупные суммы кредита увеличивают риск

👤 Молодые клиенты чаще попадают в «риск»

🕒 Длительные кредиты более проблемные

🏠 Расчётный счёт с низким балансом = высокий риск

Результаты

Пример клиента с высоким риском

- 👤 Возраст: 24 года
- 🏠 Жильё: собственное
- 💰 Сбережения: *низкий уровень*
- 🏦 Расчётный уровень: *низкий уровень*
- 💳 Сумма кредита: 9629 у.е.
- 🕒 Срок: 36 мес.
- 🎯 Цель: покупка автомобиля
- 💵 Ежемесячный платёж: 267.5




🎯 ПРОГНОЗ ДЕФОЛТА: 82%

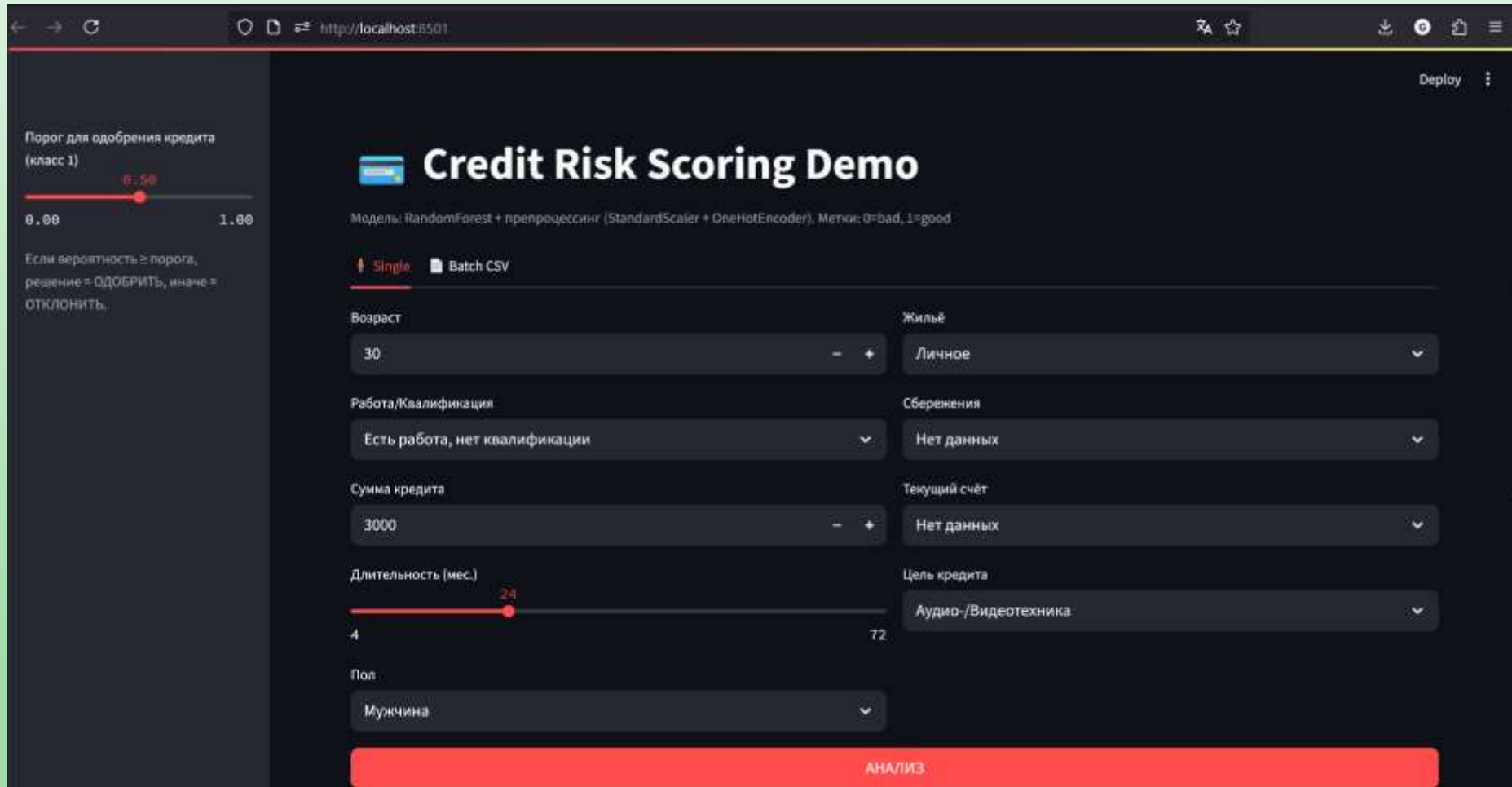
Аргумент:

- Молодой клиент, мало сбережений, большой кредит на длительный срок.

ID	Возраст	Пол	Жильё	Saving Account	Checking Account	Сумма кредита	Срок (мес.)	Цель	Ежемесячный платёж	Вероятность невозврата
812	24	male	own	little	little	9629	36	car	267.5	82%
736	23	female	rent	little	moderate	11560	24	car	481.7	76%
788	50	male	free	little	moderate	6224	48	education	129.7	77%

Практическое применение

-  **Интерактивный интерфейс (Streamlit)** - сотрудник вводит данные клиента, получает прогноз сразу.
-  **Быстрая обработка заявок** - автоматизация вместо ручной проверки.
-  **Готовность к интеграции** – модель готова к расширению.



The screenshot shows a web application titled "Credit Risk Scoring Demo" running on a local host. The interface is dark-themed and includes a sidebar on the left with a credit approval threshold slider set to 0.50. The main area contains input fields for various client attributes: Age (30), Housing (Personal), Job/Qualification (Has job, no qualification), Savings (No data), Credit Sum (3000), Current account (No data), Credit term (24 months), Credit purpose (Audio/Video equipment), and Gender (Male). A large red "АНАЛИЗ" button is at the bottom.

Порог для одобрения кредита (класс 1)

0.00 0.50 1.00

Если вероятность \geq порога, решение = ОДОБРИТЬ, иначе = ОТКЛОНИТЬ.

Credit Risk Scoring Demo

Модели: RandomForest + препроцессинг (StandardScaler + OneHotEncoder). Метки: 0=bad, 1=good

Single Batch CSV

Возраст: 30

Жильё: Личное

Работа/Квалификация: Есть работа, нет квалификации

Сбережения: Нет данных

Сумма кредита: 3000

Текущий счёт: Нет данных

Длительность (мес.): 24

Цель кредита: Аудио-/Видеотехника

Пол: Мужчина

АНАЛИЗ

Практическое применение

Возраст	20	Жильё	Аренда
Работа/Квалификация	Нет работы, нет квалификации	Сбережения	Низкий уровень
Сумма кредита	3000	Текущий счёт	Умеренный уровень
Длительность (мес.)	36	Цель кредита	Бизнес
Пол	Мужчина		

Базовое предсказание модели (argmax): **НЕНАДЁЖЕН (bad, класс 0)**

Предсказание по порогу 0.50: **НЕНАДЁЖЕН (bad, класс 0)**

Вероятности классов:

```
{
  "Вероятность дефолта (bad=0)" : 0.6233333333333333
  "Вероятность погашения (good=1)" : 0.37666666666666665
}
```

Пример РИСКОВАННОГО КЛИЕНТА:

Молодой клиент 20-ти лет запрашивает кредит в размере 3000 у.е. со сроком на 3 года на открытие/развитие бизнеса.

Проживает в арендованной недвижимости.

Уровень сбережений на низком уровне, расчётный счёт – на умеренном.

Практическое применение

Возраст	30	Жильё	Личное
Работа/Квалификация	Есть работа, есть квалификация	Сбережения	Умеренный уровень
Сумма кредита	5500	Текущий счёт	Умеренный уровень
Длительность (мес.)	24	Цель кредита	Автокредит
Пол	Женщина		

Пример НАДЁЖНОГО КЛИЕНТА:

Работающая клиент 30-ти лет открывает запрос на кредит в размере 5500 у.е. со сроком на 2 года на покупку автомобиля.

Проживает в личной недвижимости. Уровень сбережений и расчётного счёта - умеренный.

Базовое предсказание модели (argmax): НАДЁЖЕН (good, класс 1)

Предсказание по порогу 0.50: НАДЁЖЕН (good, класс 1)

Вероятности классов:

```
{
  "Вероятность дефолта (bad=0)" : 0.23666666666666666
  "Вероятность погашения (good=1)" : 0.7633333333333333
}
```