

# Анализ кредитных рисков: факторы дефолтов заемщиков

На основе данных кредитных заявок

Петрова Е. Д.



# Введение

## Цель:

Выявить ключевые факторы, влияющие на вероятность дефолта заемщиков, и сегментировать группы риска.

## Гипотезы:

- 1) Дефолты связаны с низким доходом и высокими кредитными нагрузками.
- 2) Демография (возраст, семейное положение) влияет на платежную дисциплину.



## Описание данных

Этот набор данных содержит 3 файла:

1. 'application\_data.csv' содержит всю информацию о клиенте на момент подачи заявки. Данные о том, есть ли у клиента трудности с оплатой.
2. 'previous\_application.csv' содержит информацию о предыдущих данных о кредите клиента. Он содержит данные о том, была ли предыдущая заявка одобрена, отменена, отклонена или неиспользована.
3. 'columns\_description.csv' — словарь данных, который описывает значение переменных.

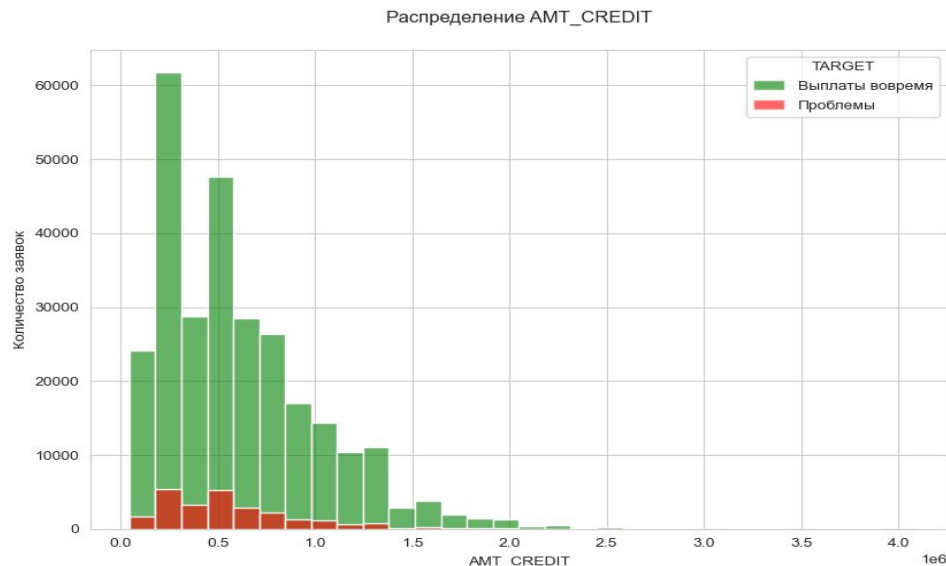
## Анализ целевой переменной



Сильный дисбаланс классов: 92% заемщиков исправно платят, 8% — дефолт.

# Анализ числовых переменных

## Основные финансовые показатели



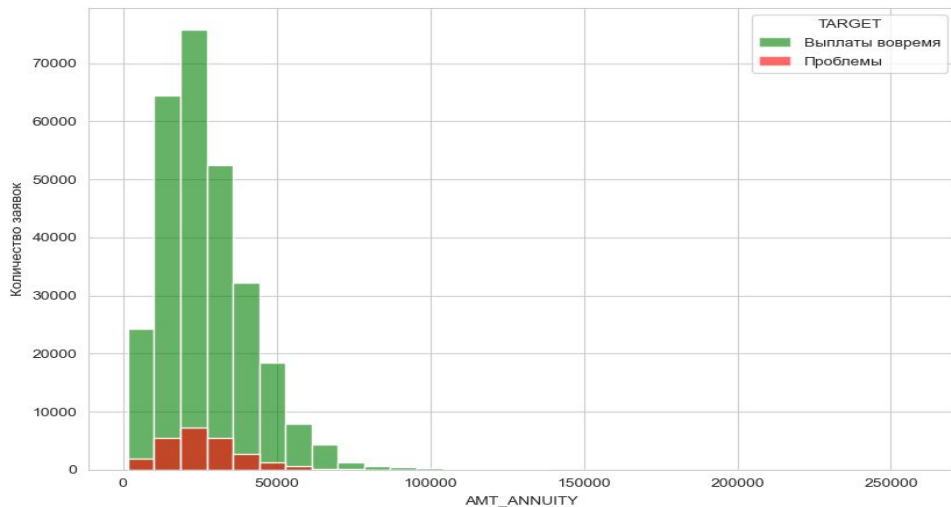
Пик дефолтов при 200–500 тыс.  
— заемщики с низким доходом  
не справляются с нагрузкой.

Крупные кредиты (>1 млн)  
выплачиваются стабильно  
(высокий доход заемщиков)

# Анализ числовых переменных

## Основные финансовые показатели

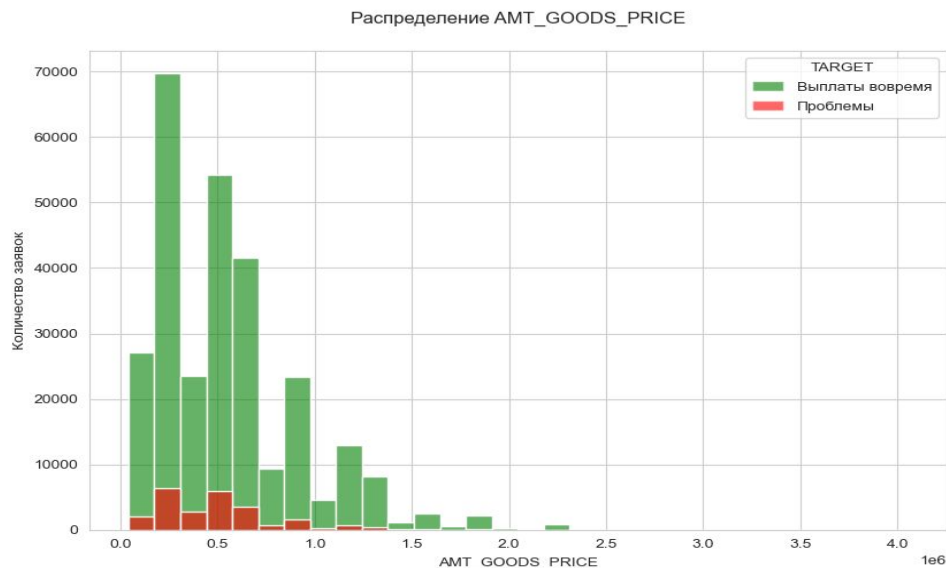
Распределение AMT\_ANNUITY



Риск дефолта максимален при ежемесячных платежах ~25 тыс./мес.

# Анализ числовых переменных

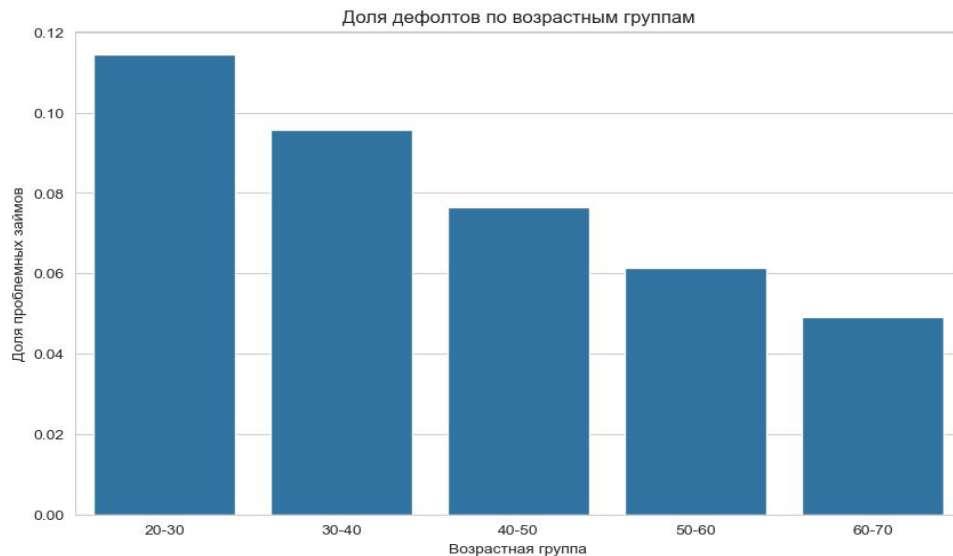
## Основные финансовые показатели



Пик дефолтов при 200–500 тыс. — заемщики с низким доходом не справляются с нагрузкой потребительского кредита

# Анализ числовых переменных

## Демографические данные

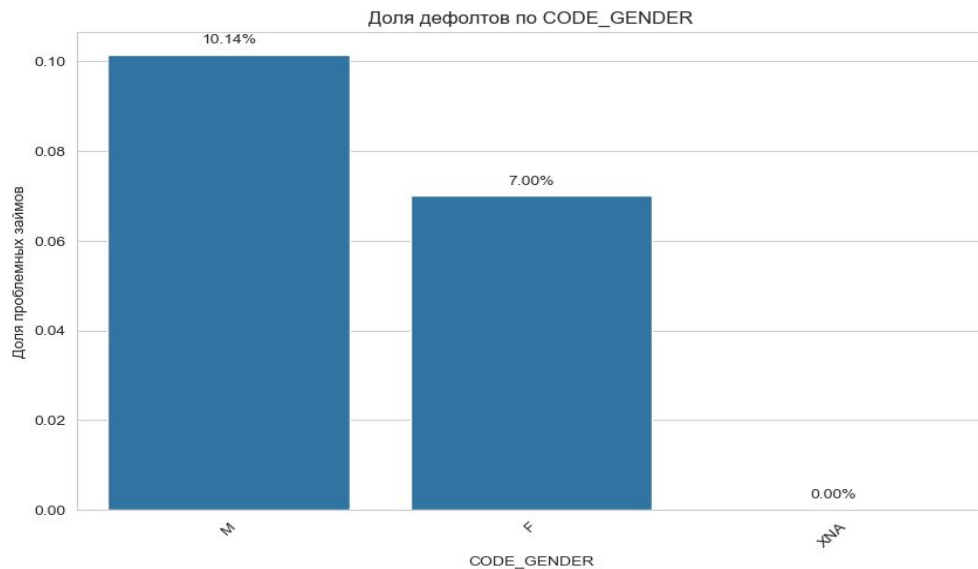


Группа риска — 20–40 лет  
(пик кредитной активности).  
После 40 лет риск постепенно  
снижается.



# Анализ категориальных переменных

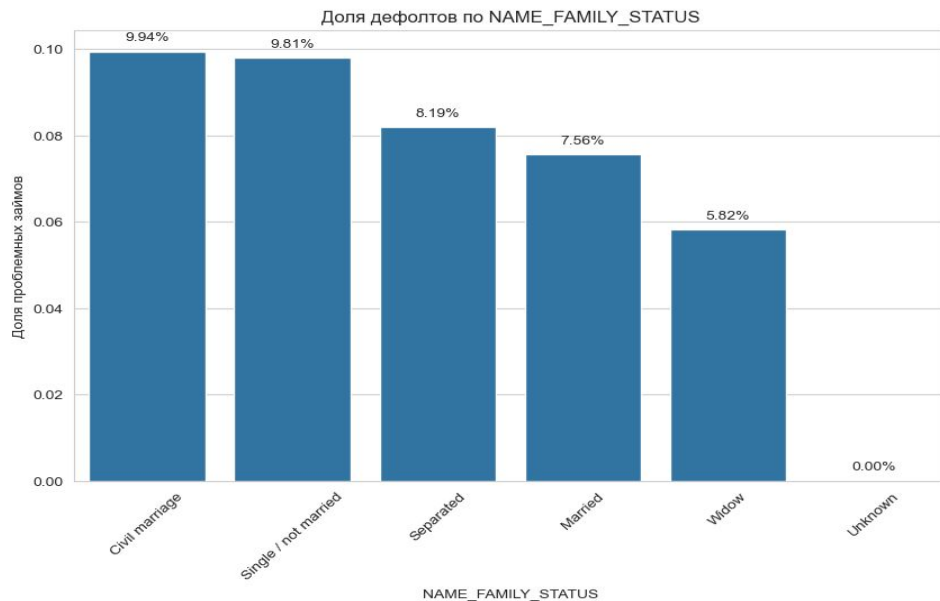
## Демографические данные



Мужчины чаще допускают дефолт (10.1% vs 7% у женщин)

# Анализ категориальных переменных

## Демографические данные

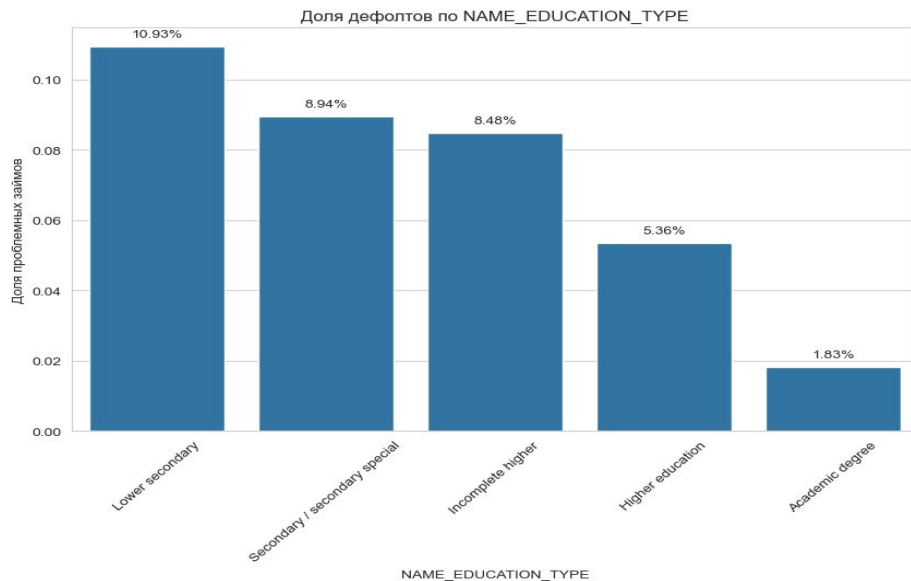


Наибольший риск у незамужних и гражданских партнеров.

Минимальный риск у вдовцов.

# Анализ категориальных переменных

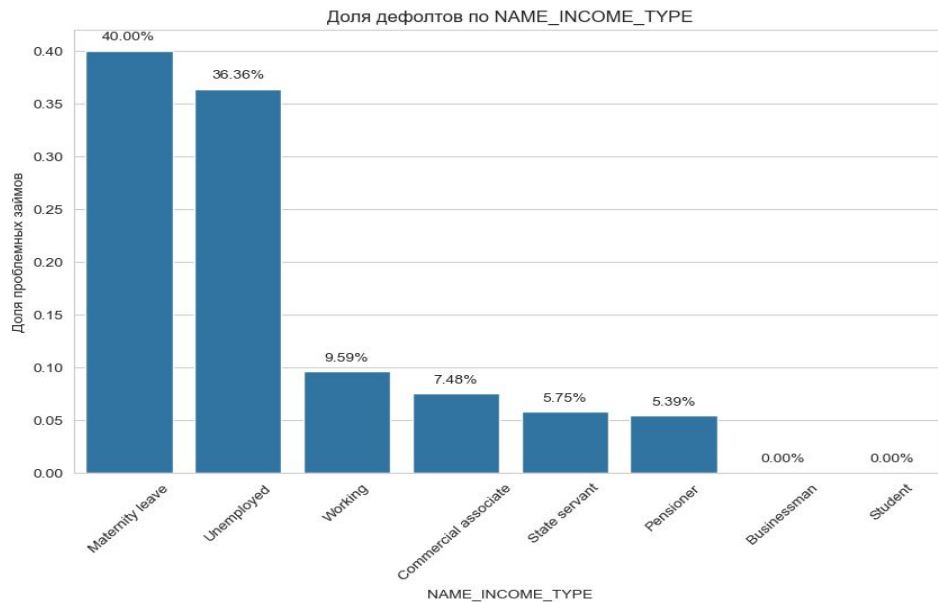
## Демографические данные



Образование → прямая зависимость (чем меньше степень, тем вероятность дефолта выше)

# Анализ категориальных переменных

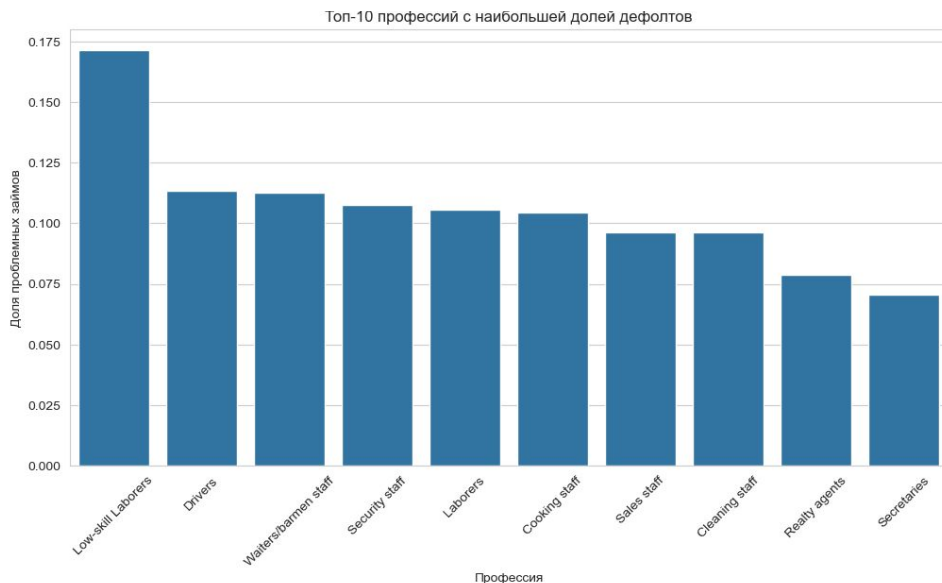
## Финансы и занятость



Безработные и в декрете →  
высокий риск  
Бизнесмены/студенты → почти  
нет дефолтов

# Анализ категориальных переменных

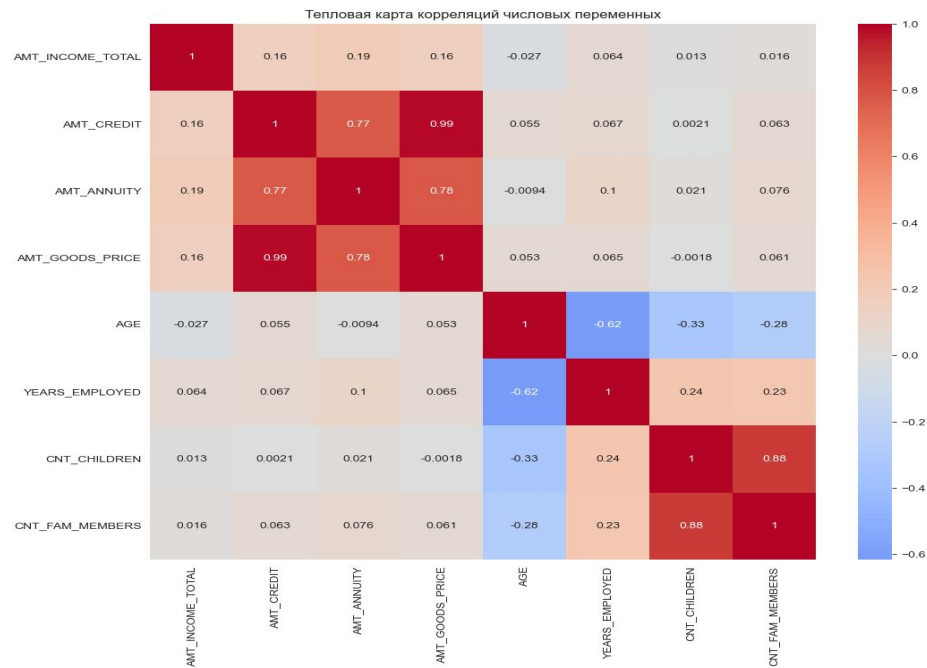
## Финансы и занятость



Топ-10 профессий  
с наибольшей долей  
дефолтов

Рабочие профессии → группа  
риска

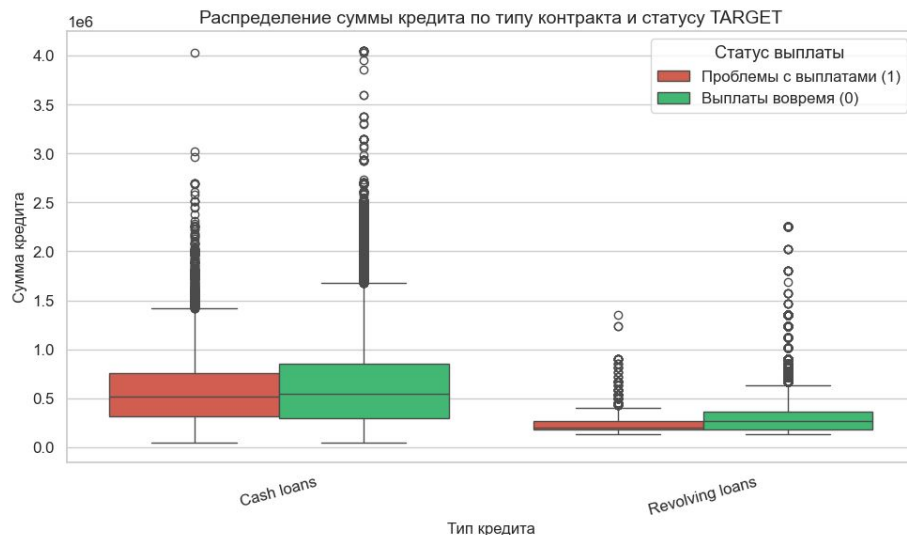
# Анализ корреляций



- 1) Сумма кредита почти полностью зависит от стоимости товара
- 2) Чем дороже товар, тем выше аннуитет
- 3) Размер кредита сильно влияет на размер аннуитета
- 4) Чем старше человек, тем меньше детей — дети взрослеют и выходят из состава домохозяйства
- 5) Возраст имеет умеренно отрицательную связь с трудовым стажем — возможно, требует дополнительного анализа или очистки данных (например, стаж может быть обрезан по текущему работодателю)
- 6) Доход почти не влияет на сумму кредита или товара

# Бивариативный анализ

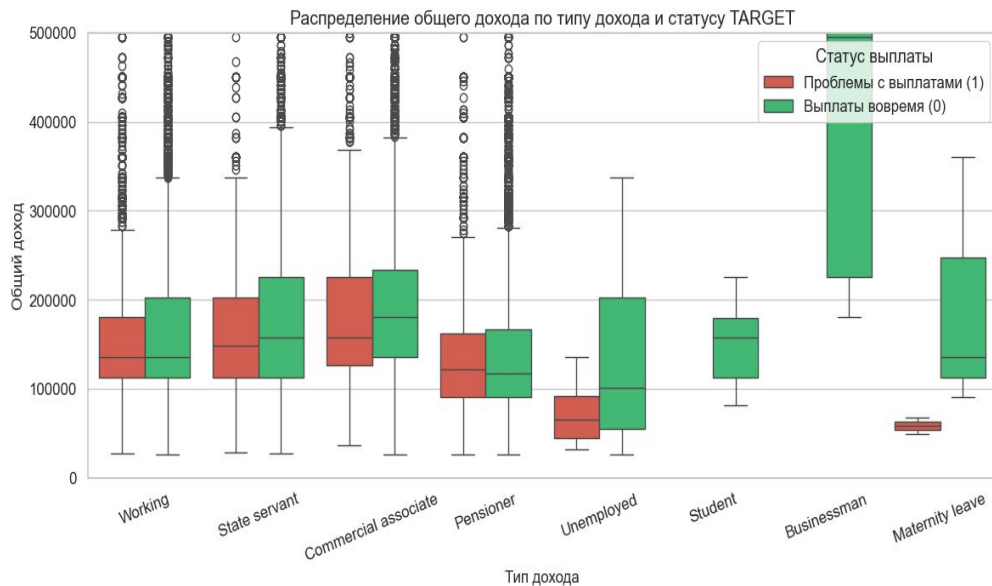
## Числовые vs Категориальные



- Много выбросов в обеих категориях контрактов
  - В пределах каждого типа контракта медиана суммы кредита не сильно отличается между теми, кто платил вовремя и теми, кто допускал просрочки.
  - Кредитные суммы при потребительских кредитах (Cash loans) варьируются шире, чем в Revolving loans (например, кредитные карты).
- Это означает, что сама сумма кредита не является ключевым фактором риска, по крайней мере в разрезе контрактного типа

# Бивариативный анализ

## Числовые vs Категориальные



Выбросов почти нет у категорий:

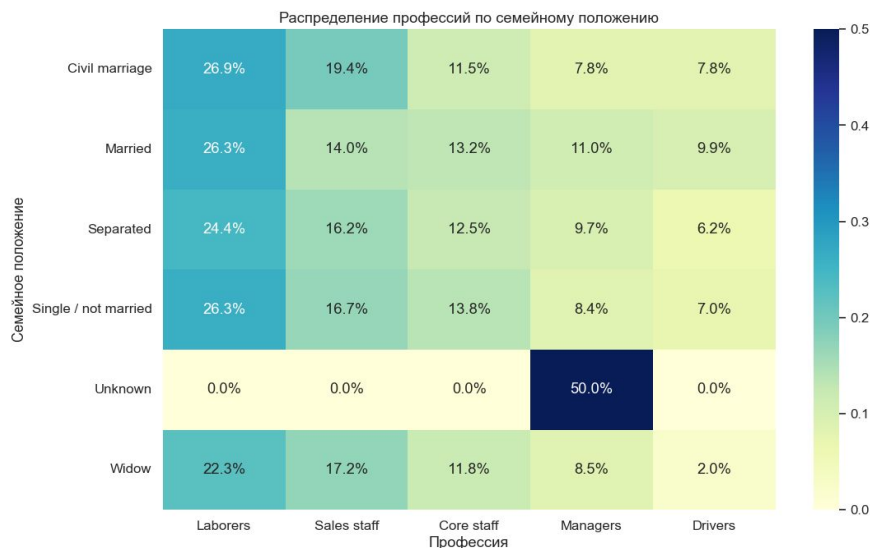
- Безработные
- Студенты
- ИП / предприниматели (Businessman)
- В декрете (Maternity leave)

Это может быть связано с: малой выборкой в этих группах, фиксированными доходами (например, студенты, декрет)



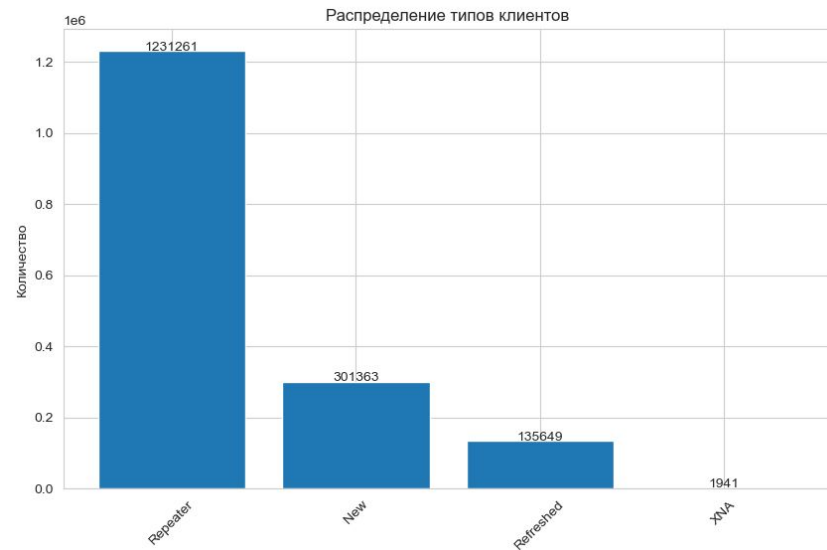
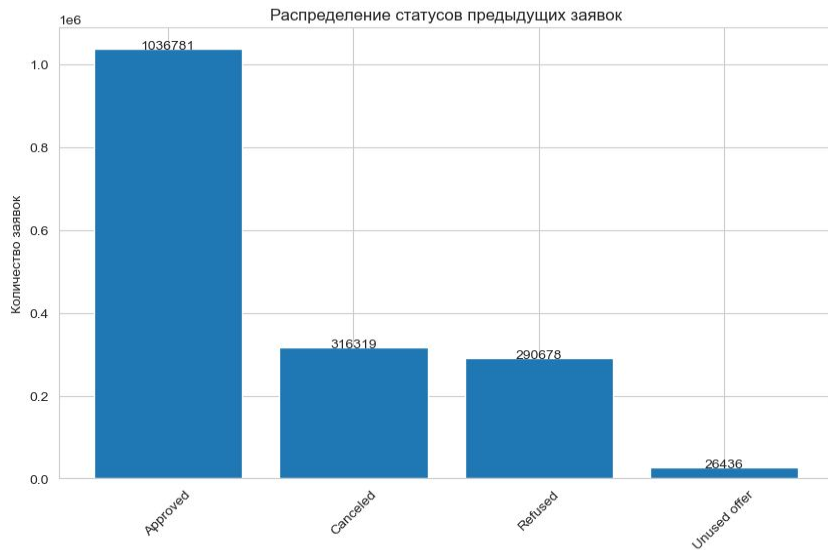
# Бивариативный анализ

## Категориальные vs Категориальные



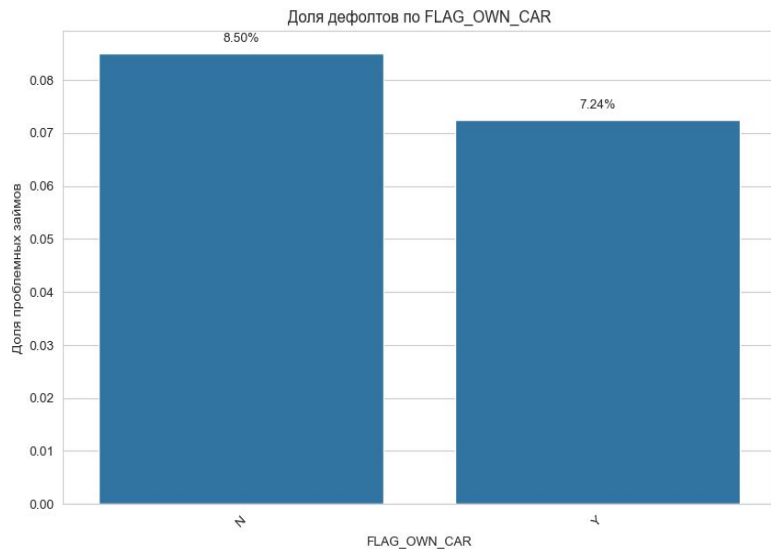
- 1) Рабочие — наиболее массовая категория заёмщиков, независимо от семейного статуса
- 2) Менеджеры — более распространены среди официально женатых
- 3) Sales staff и Core staff - стабильно вторые по частоте во всех группах
- 4) Drivers (водители) — реже встречаются среди вдов и разведённых
- 5) Категория "Unknown" — аномалия 50% — менеджеры, остальные профессии отсутствуют

# Анализ предыдущих заявок



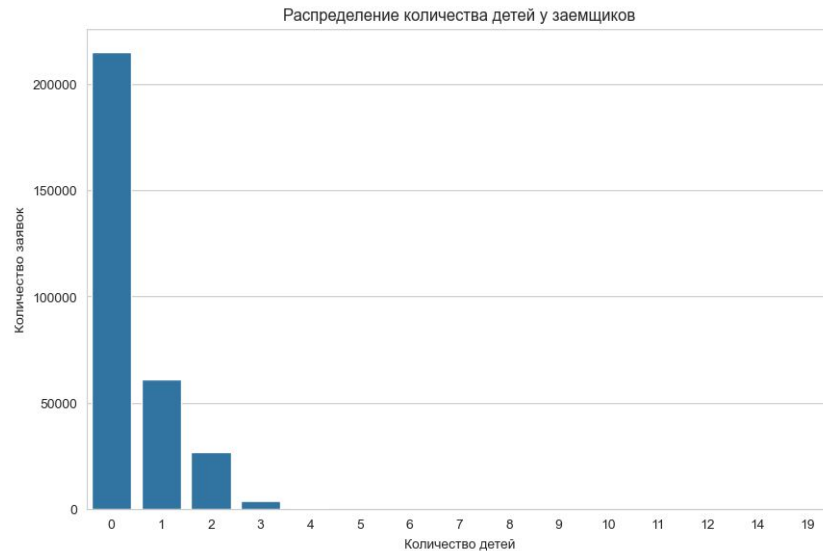
Больше всего хотят получить кредит, те кто уже были одобрены и те кто уже брали.

## Дополнительный анализ



Нет машины → выше риск (8.5% vs 7.24%)

Машина → признак стабильности



Большинство без детей или 1 ребенок

Многодетные (3+) редко берут кредиты



# Предобработка данных: ключевые этапы

**Обработка выбросов** — ограничение экстремальных значений с помощью медианы и MAD

**Обработка предыдущих заявок** (previous\_application):

- Абсолютные значения дней решения
- Вычисление отношения суммы заявки к кредиту
- Создание бинарного признака одобрения заявки


**Агрегация признаков по клиенту:** средние, максимальные, суммарные значения, количество заявок, частота одобрений, временные метрики

**Объединение агрегированных данных с основным датасетом**

**Инженерия новых признаков:**

- Отношения кредитов и дохода
- Отношение трудового стажа к возрасту
- Возраст, количество дней между заявками
- Отношение числа детей к членам семьи
- Комбинация внешних источников риска (EXT\_SOURCE\_COMBINED)

**Финальная обработка выбросов по новым признакам**  
**Удаление лишних столбцов и заполнение пропусков**



# Логистическая регрессия

- ROC-AUC: 0.728 | F1-score: 0.279
- Точность (класс 0): 94% | Полнота (класс 1): 43%
- Ошибки: 8,206 FP | 2,833 FN

Результаты:

- Выявляет 43% проблемных клиентов (класс 1)
- Точность предсказаний "хороших" клиентов (класс 0) — 94%
- Общая точность (accuracy) — 82%




# Balanced Bagging + Logistic Regression

- ROC-AUC: 0.728 | F1-score: 0.278
- Точность (класс 0): 94% | Полнота (класс 1): 42%
- Ошибки: 7,821 FP | 2,901 FN

Результаты:

- Почти идентична обычной LogReg (F1-score 0.278 vs 0.2786)
- Чуть лучше accuracy (83%)



## Модель: CatBoost

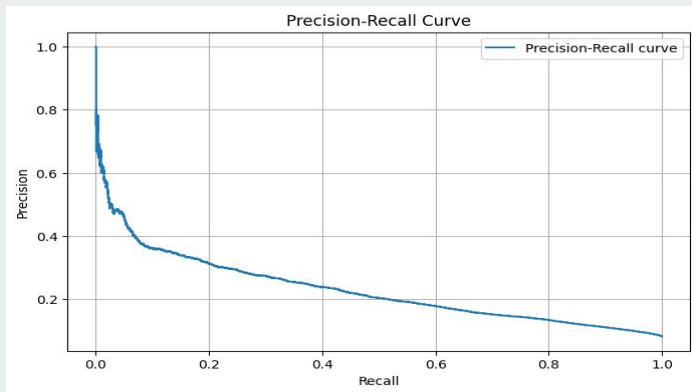
- ROC-AUC: 0.741 | F1-score: 0.292
- Точность (класс 0): 95% | Полнота (класс 1): 44%
- Ошибки: 7,747 FP | 2,796 FN

Результаты:

- Лучший recall (44%) для класса 1
- Высокая точность для класса 0 (95%)
- Порог срабатывания всего 0.11 (очень "чувствительная" модель)

Лучший выбор, если критично не пропустить проблемных клиентов

# Модель: XGBoost



- ROC-AUC: 0.744 | F1-score: 0.298
- Точность (класс 0): 94% | Полнота (класс 1): 38%
- Ошибки: 5,923 FP | 3,05 FN

Результаты:

- Лучший F1-score (0.298) для класса 1
- Максимальный ROC-AUC (0.744)
- Высокая общая точность (85%)

Выявляет меньше рискованных клиентов (recall 38%), чем CatBoost

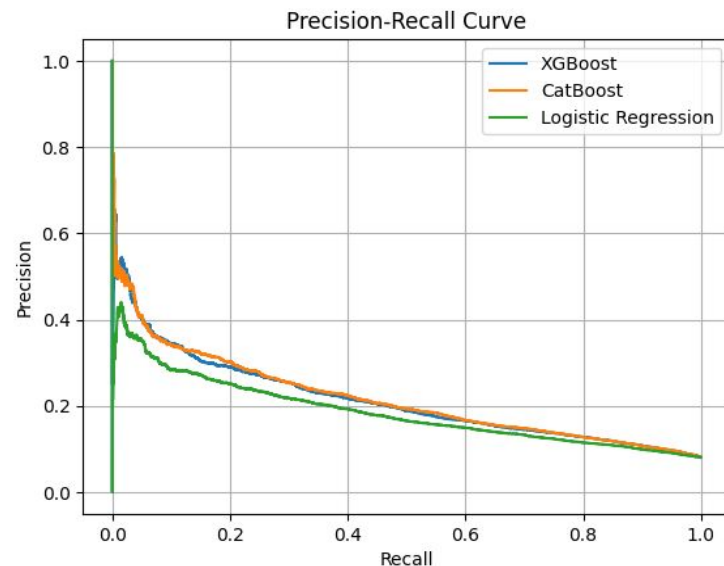


## Результаты моделей с SMOTE

Модель	ROC-AUC	PR-AUC	F1-score
XGBoost	0.7332	0.2125	0.2829
CatBoost	0.7356	0.2160	0.2875
LogReg	0.7029	0.1846	0.2605

По сравнению с моделями без SMOTE:

- Небольшое снижение всех ключевых метрик (ROC-AUC, PR-AUC, F1).
- Повышение recall у минорного класса (1), но за счёт просадки precision.
- Особенно страдает логистическая регрессия, что типично при переобучении на синтетических данных.





## Общие выводы

1. Если цель — баланс метрик: XGBoost (F1-score 0.298, ROC-AUC 0.744)
2. Если цель — минимизация риска: CatBoost (recall 44% против 38% у XGBoost)
3. Для интерпретируемости: Logistic Regression



**Спасибо за внимание!**