

Предсказание отмены брони



Что есть в наших данных о госте

Отмена брони

Отменял ли бронь гость до этого

Время

Дата заезда и отъезда

Количество гостей

Человек приедет с семьей , детьми или один

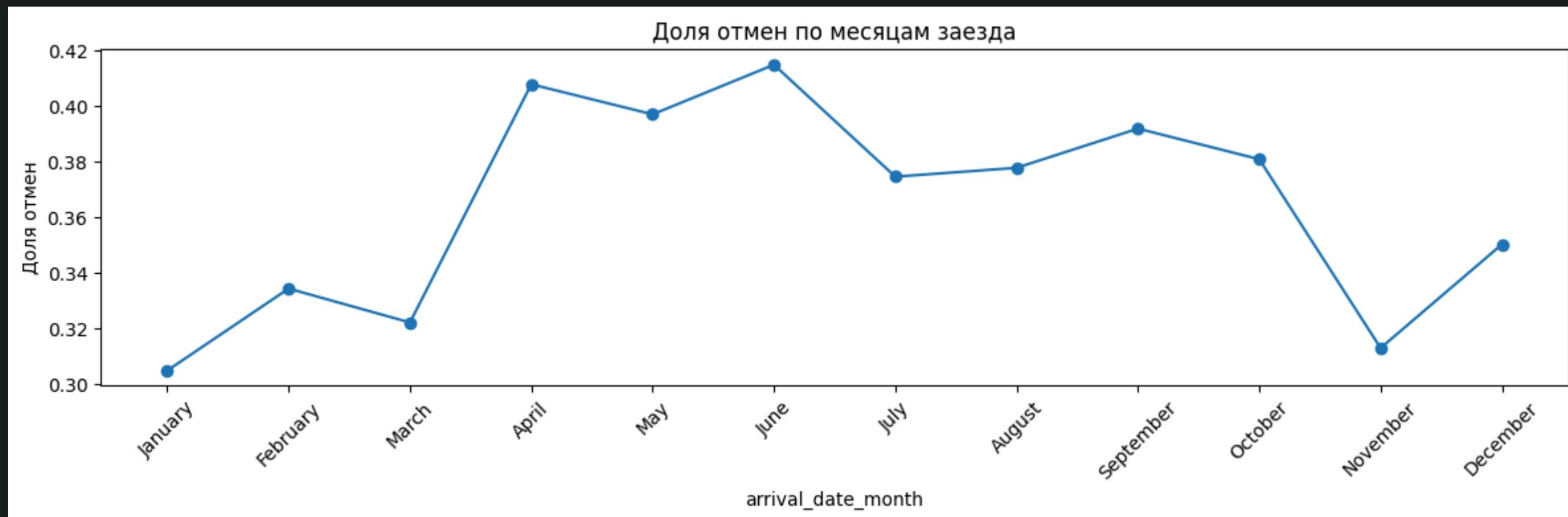
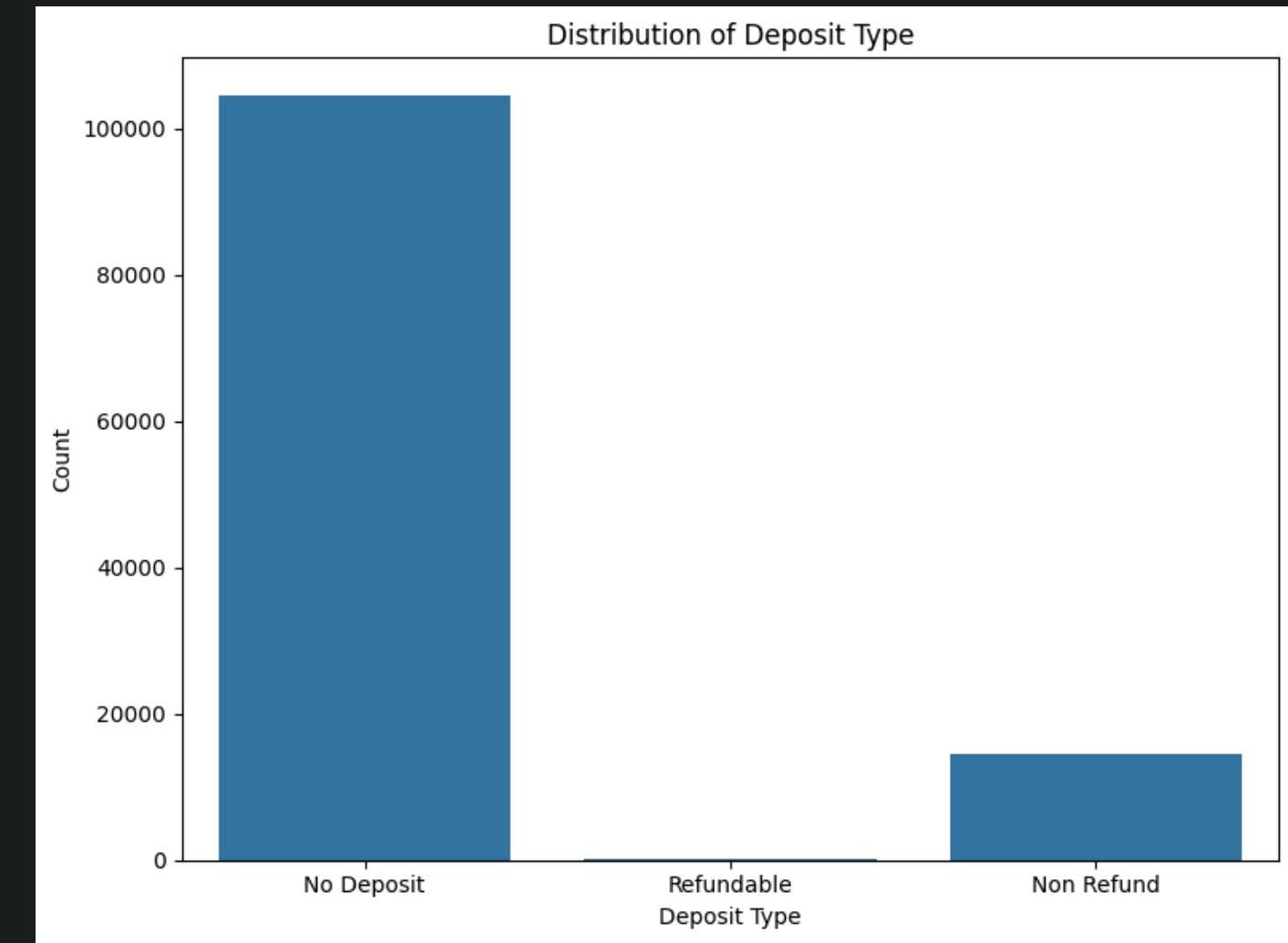
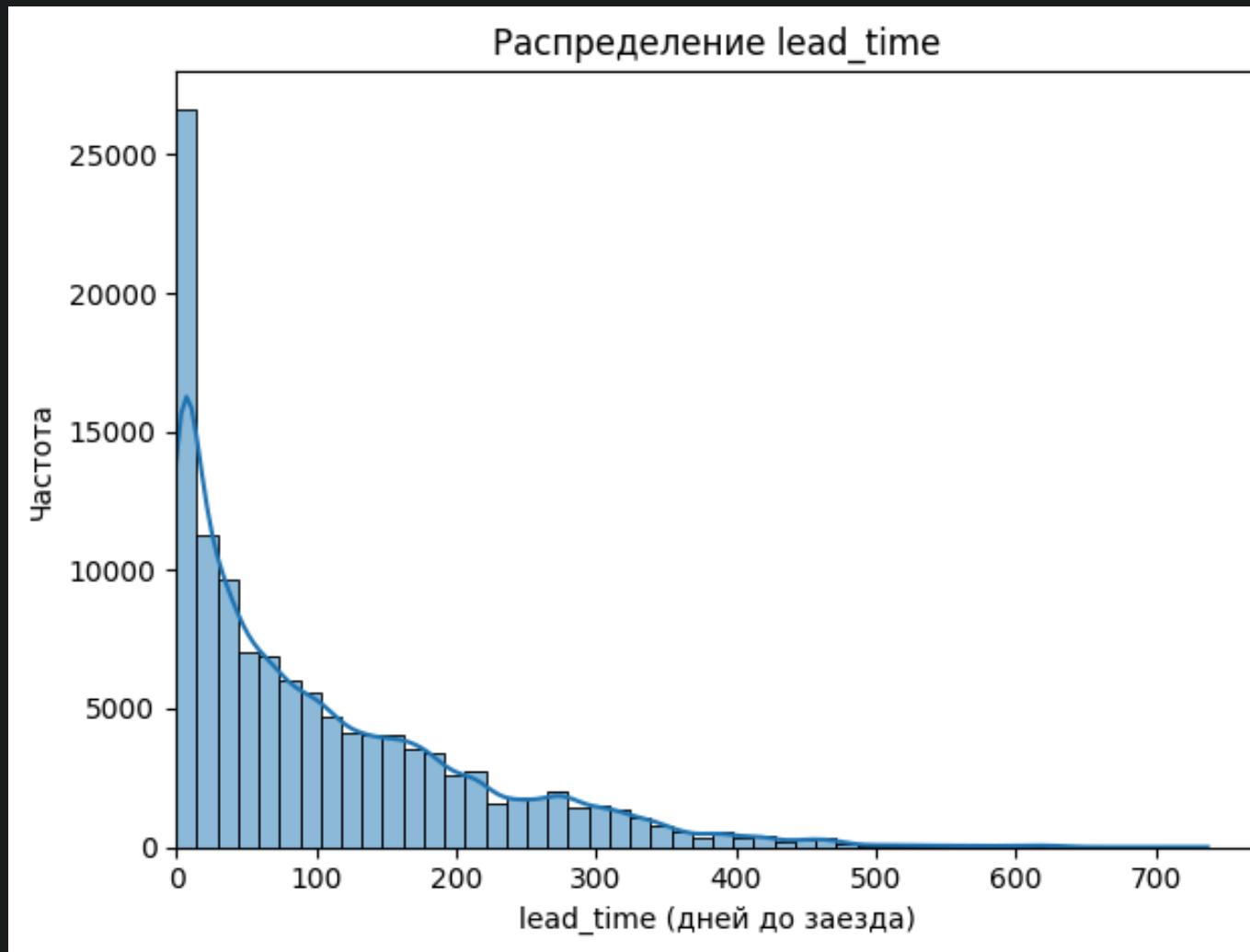
Сколько дней до визита

Насколько заранее гость сделал бронирование

Предобработка данных

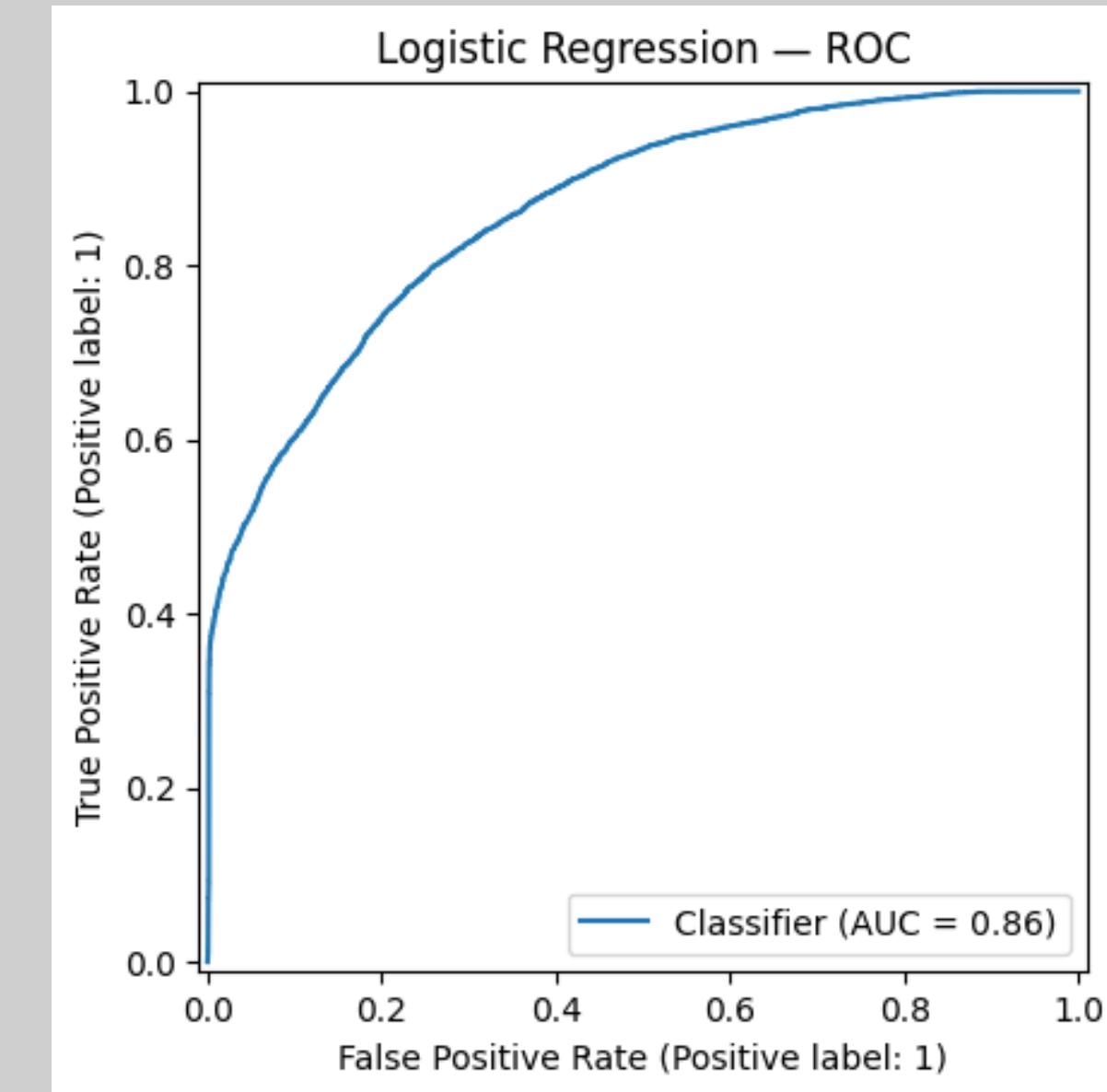
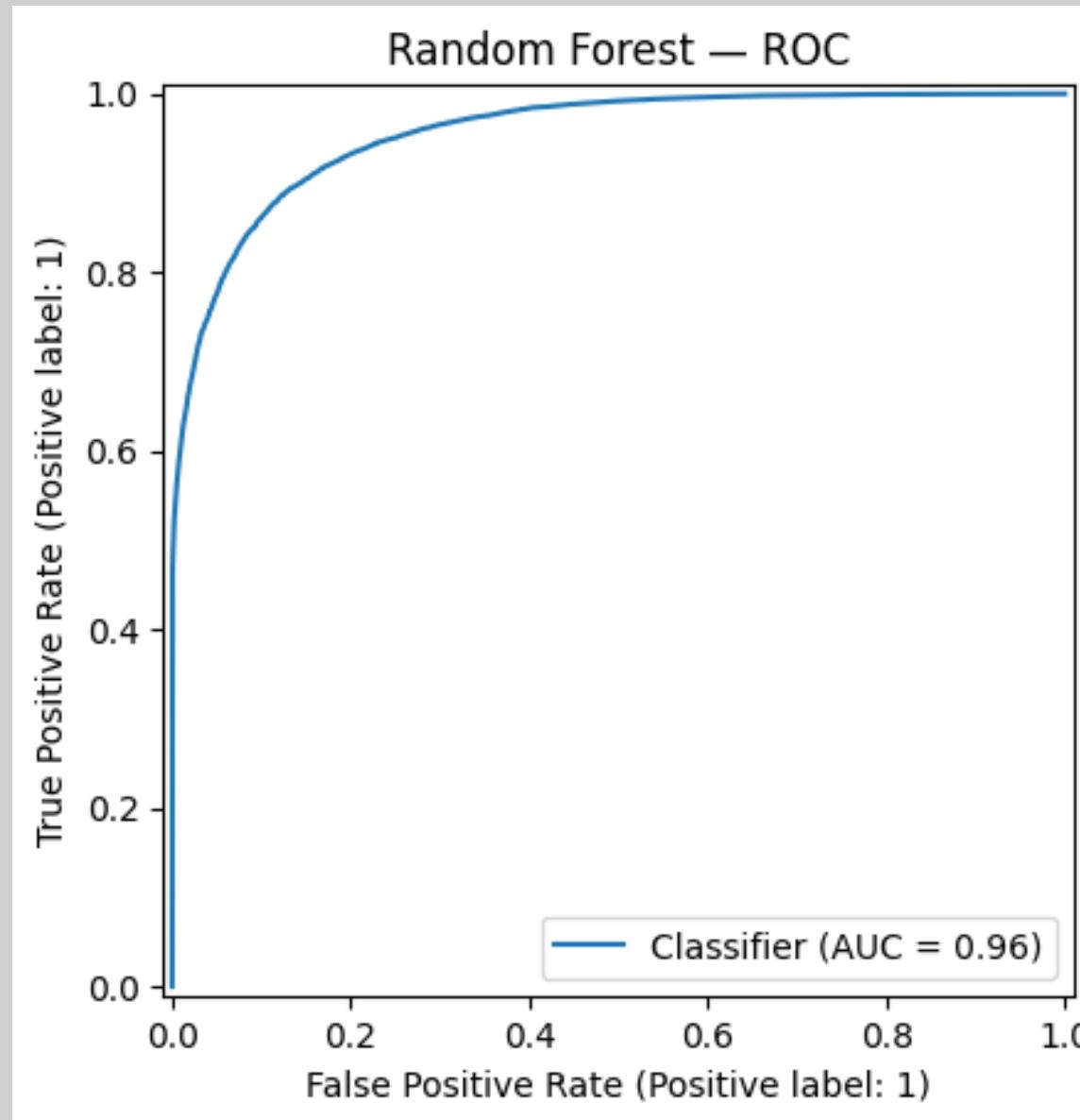
- 01 Удалили столбцы
'agent','company'
- 02 Удалили тех , кто указал
0-ое количество гостей
- 03 Заполнили пропуски в
столбце 'children' модой
- 04 Удалили столбцы, которые
явно дают утечку -
'reservation_status',
'reservation status date'

Преданализ



Бейзлайн

Построили логистическую регрессию и
случайный лес



Работа с аномалиями

Методы поиска аномалий

Статистические методы

Через IQR, Z-оценку и тест Граббса

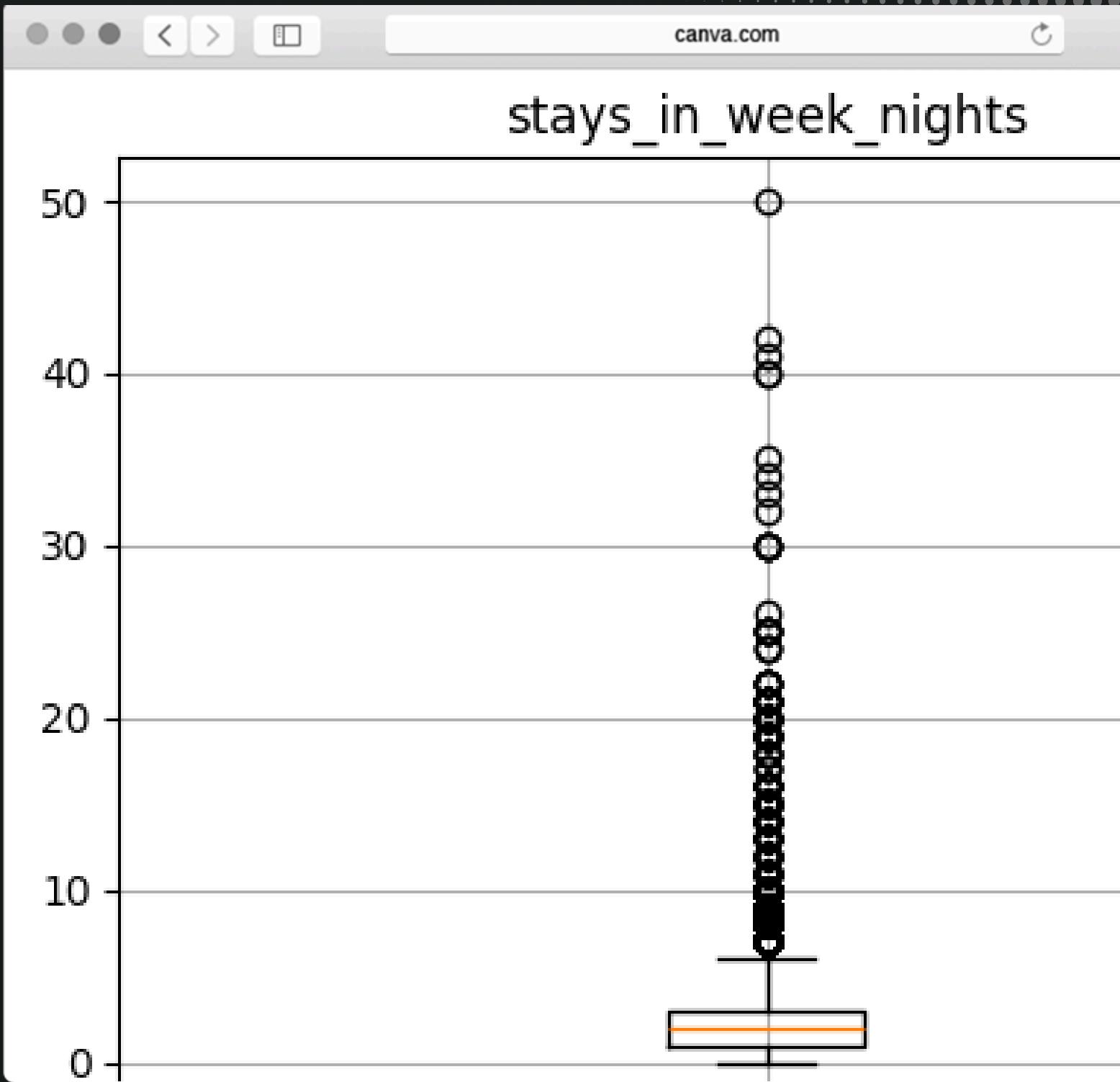
ML-методы

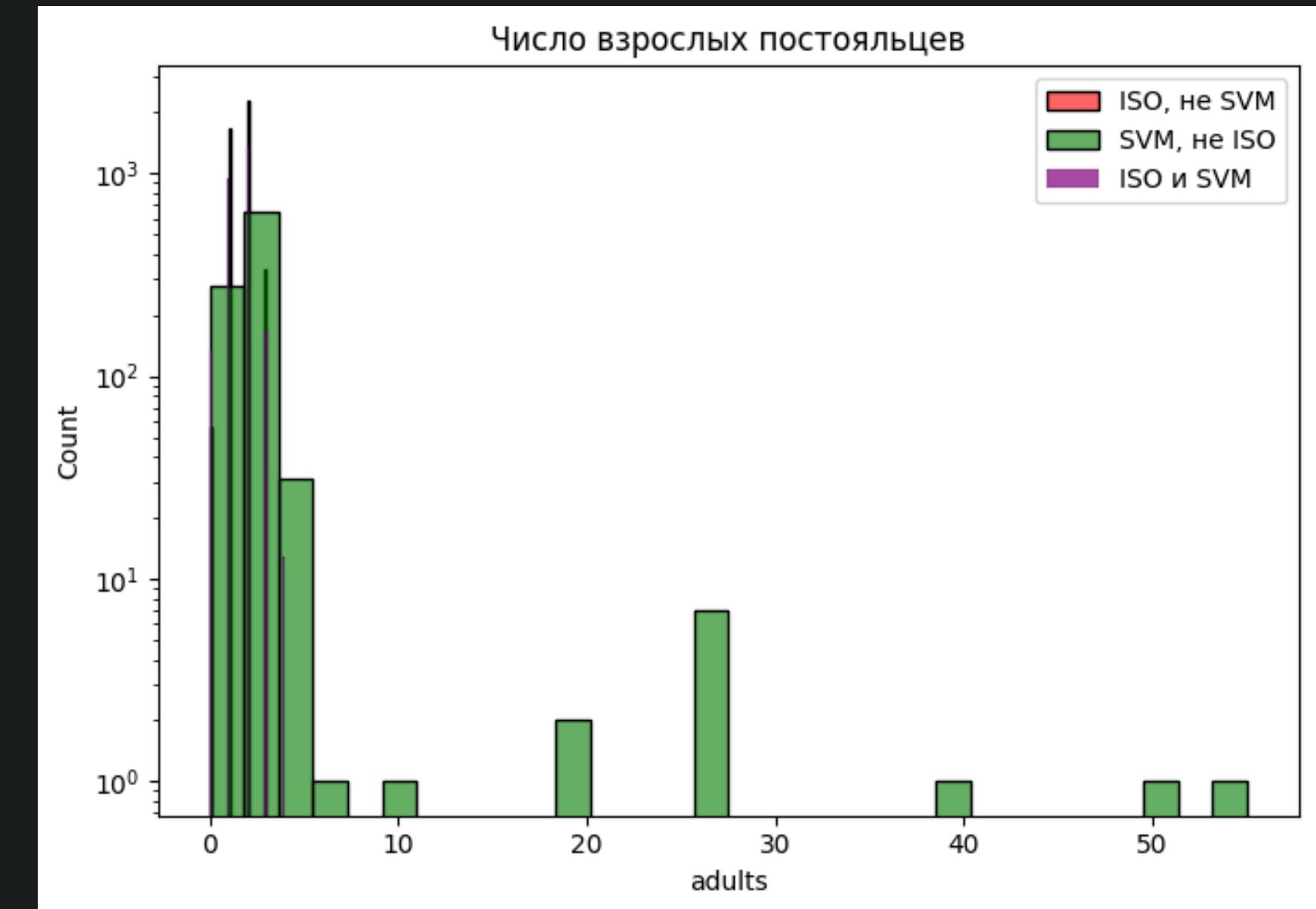
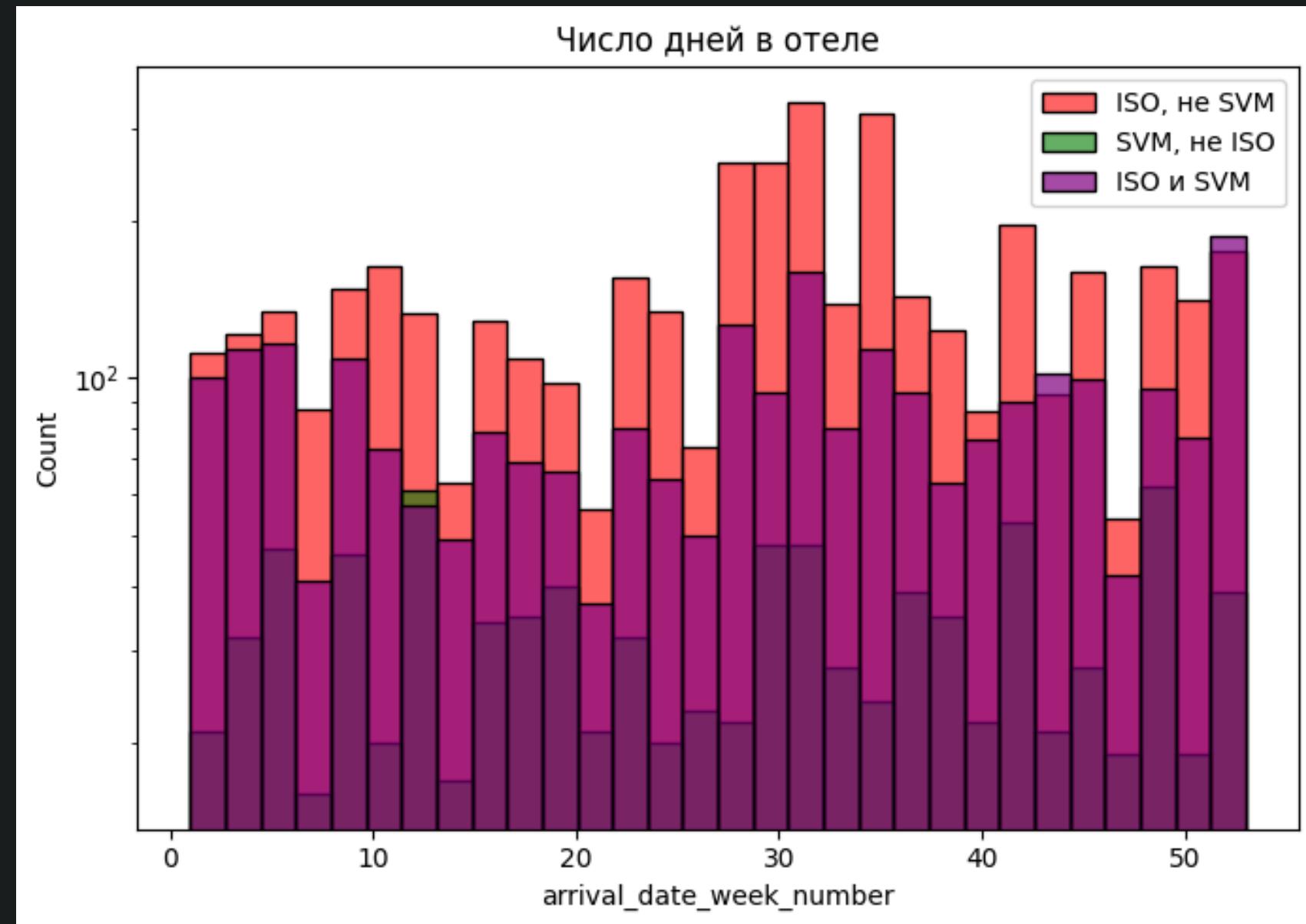
Использовали IsolationForest и
OneClassSVM

Где были найдены аномалии

- Очень раннее бронирование номера
- Бронирование номера на нетипично долгое время
- Необычно большое число заселяющихся
- Большое число необычных запросов клиента

Эти строки отражают аномалии по распределению и бизнес-контексту





А что не является аномалиями

Большая средняя
ставка за ночь



Зависит от числа
постояльцев

Большое изм.
числа статусов
бронирования



Зависит от частоты
бронирования
номеров

Генерация признаков и отбор переменных



Процессы отбора и генерации фич

- **Отбор фичей**
Удалили коррелирующие и малоинформативные признаки
- **Временные признаки**
Добавили цикличности временным признакам
- **Генерация новых фич**
Создали агрегаты, удельные метрики и нормализовали время до заезда.
- **Генерация на основе соседей**
За счет координат сгенерировали новые фичи на основе близости с соседями
- **Обработка категориальных переменных**
Применение TargetEncoding

Детальное описание генерации фич

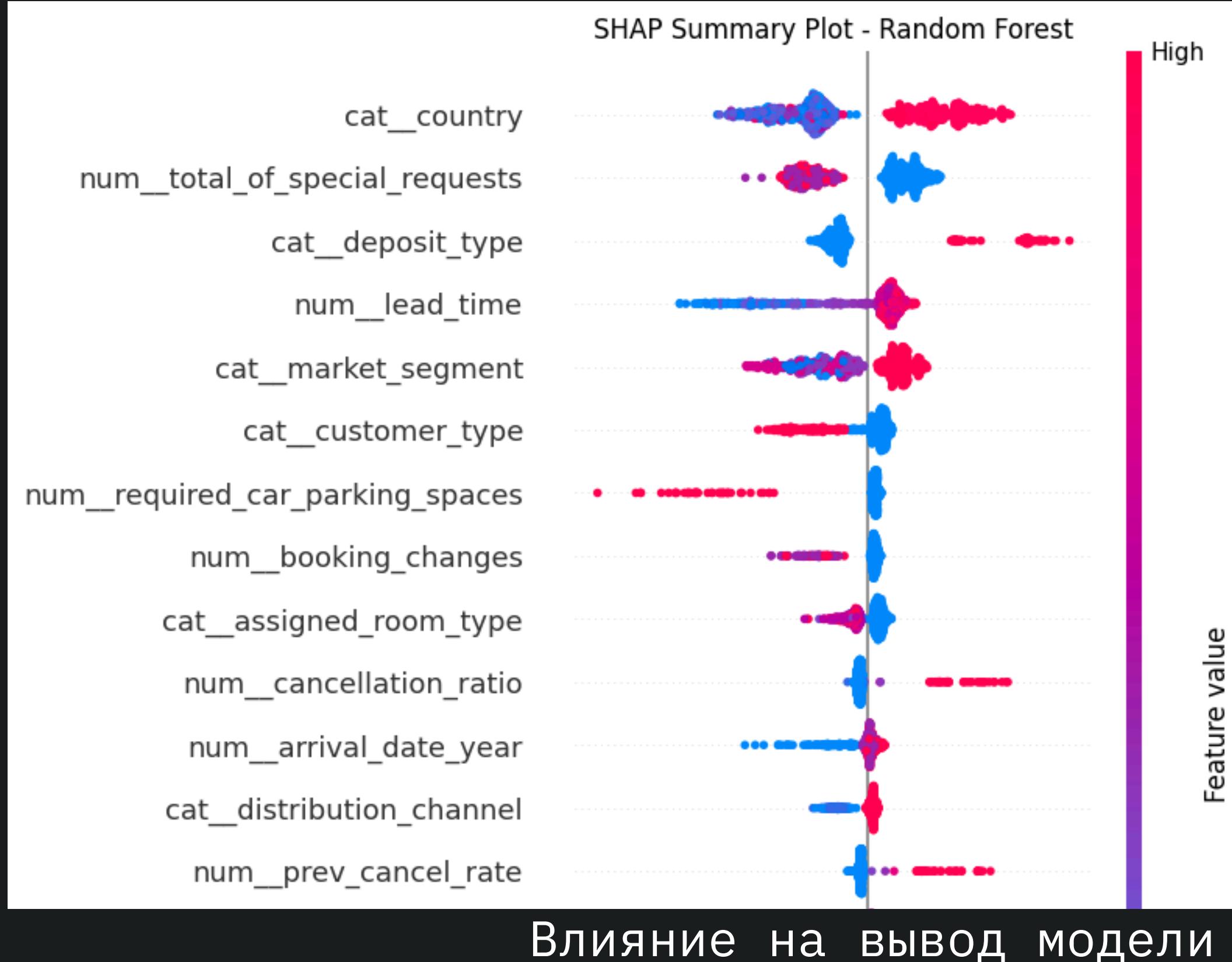


Отбор фичей	Временные признаки	Генерация новых фич	Генерация новых фич на основе соседей	Обработка категориальных признаков
По результатам EDA удалили неинформативные признаки	Добавили цикличность всем временных признакам Трансформация фич по инсайтам из EDA	Добавили суммарное представление о количестве проведенных дней Информация об общем количестве людей и бинарный признак, являются ли они семьей Перерасчет части характеристик в относительные величины	Сгруппировали фичи на 3 группы - о жильцах, заказе и брони По этим данным обучили модель для определения принадлежности каждой группе	LabelEncoding категориальных фичей Добавление TargetEncoding для типа депозита

Интерпретация и диагностика моделей

с помощью `shap` и `lime`

Shap для обеих моделей



Страна, запросы и тип депозита – главные факторы риска

Топ 10 фичей из наших моделей

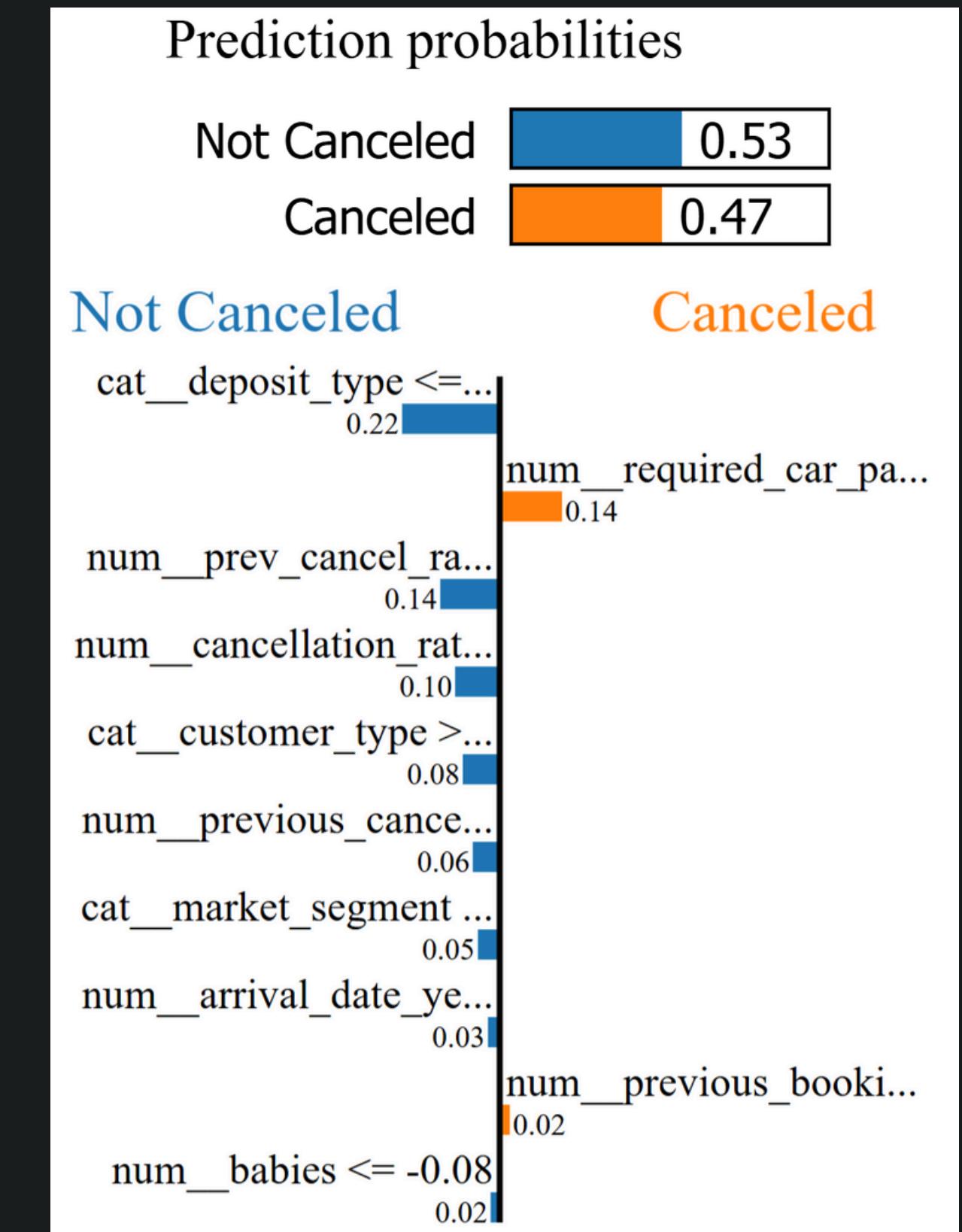
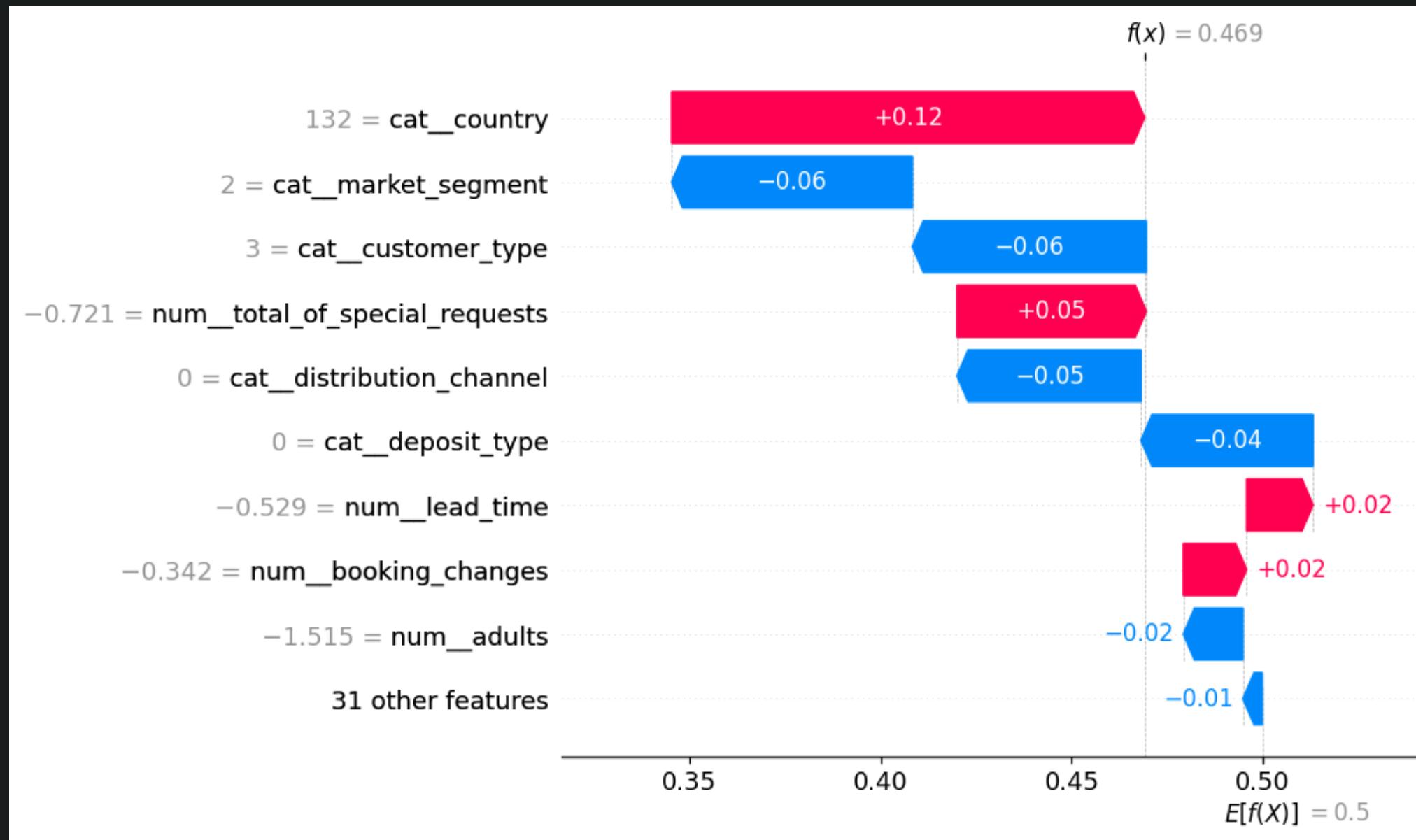
cat__assigned_room_type	0.842243
cat__deposit_type	0.787011
cat__reserved_room_type	0.682291
cat__market_segment	0.641505
num_required_car_parking_spaces	0.593236
cat__country	0.537620
num__cancellation_ratio	0.492930
num__lead_time	0.478895
num_total_of_special_requests	0.441425
cat__distribution_channel	0.304214

Logistic Regression

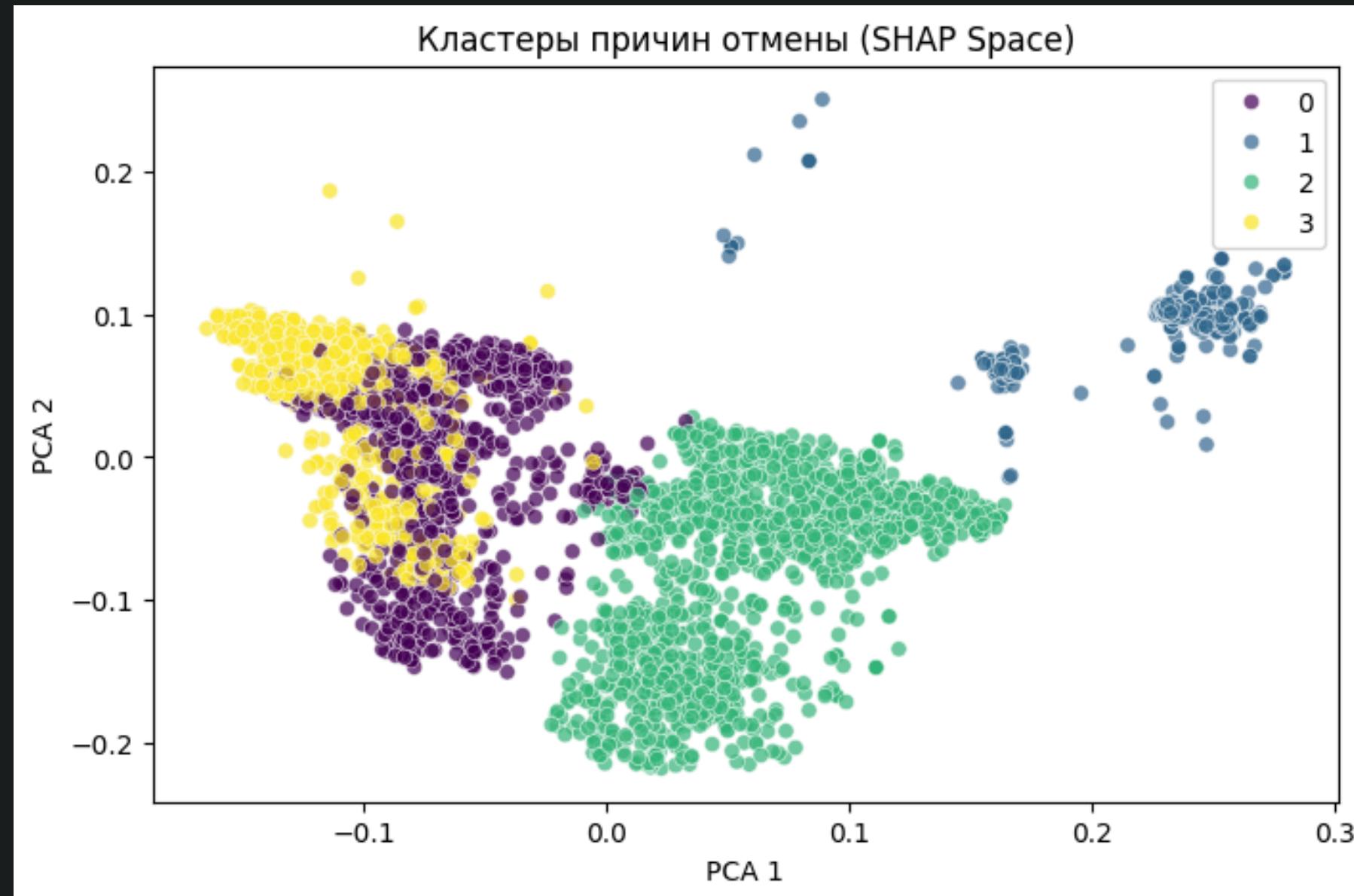
cat__country	0.095199
num__total_of_special_requests	0.064048
cat__deposit_type	0.058635
num__lead_time	0.055919
cat__market_segment	0.054408
cat__customer_type	0.030006
num__required_car_parking_spaces	0.022052
num__booking_changes	0.020369
cat__assigned_room_type	0.018128
num__cancellation_ratio	0.015972

Random Forest

Lime и shap на 1 наблюдении



Кластеры причин отмены



Кластеризация выделила 4
устойчивых паттерна поведения

Влияние кластеризации
F1 (Без кластеров): 0.7411
F1 (С фичей кластера): 0.7583
Прирост: +0.0172

Топ 5 фичей с НЕОБЫЧНОЙ связью

Feature	Correlation
num__total_of_special_requests	-0.832155
cat__assigned_room_type	-0.811254
cat__arrival_date_month	-0.805187
num__booking_changes	-0.624865
cat__customer_type	-0.564639

Топ 4 признака с ПРЯМОЙ связью

Feature	Correlation
cat__market_segment	0.799109
num__stays_in_week_nights	0.803557
num__total_nights	0.879020
num__adr	0.919236

Shapley Flow

Анализ аномалий (Shapley Flow)

Найдено аномалий поведения: 250 (из 5000)

F1 Score после Shapley Flow очистки: 0.7886

Для сравнения (SHAP-only без очистки): 0.8178

Для сравнения (Original Baseline): 0.7663

Аномалии в SHAP-пространстве оказались не шумом, а важными редкими кейсами

Интерпретация кластеров Shapley Flow (Топ отличий):

Flow_Cluster	cat__country	cat__deposit_type	cat__market_segment	num__total_of_special_requests	num__lead_time
0	-0.065288	-0.033865	0.050625	-0.032257	0.000461
1	0.113555	0.192212	-0.024577	0.055075	0.030405
2	0.095879	-0.048548	-0.019330	0.001671	-0.031369
3	-0.122326	-0.037815	-0.088792	0.010743	-0.002557

Спасибо !

