Прогноз преступности в районах Нью-Йорка

Hackathon project

Авторы: Малышева Анастасия, Петрухин Владислав, Картвелишвили Евгений, Сидоренко Николай

Проблема

Большинство туристов хотят себя чувствовать себя в безопасности в новых для них местах. Однако, во всех городах без исключения пребывающие имеют риск столкнуться с преступностью.

Цель проекта: обучить модель машинного обучения для туристов, которая будет предсказывать приезжим горожанам наиболее преступные районы в выбранные месяцы/дни/часы по местоположению и типам мест (например, станция, парк и т.п).

Данные

В проекте использован датасет **«New York City Crimes»** (источник: NYC OpenData)

Описание: в датасете содержатся отчеты о преступлениях в районах Нью-Йорка с 1900 по 2022 год с информацией о типе преступления, времени, жертве, месте и иных характеристиках.

Детали: датасет содержит 35 уникальных столбцов и около 8 миллионов строк.

Обработка датасета

1

Удаление ненужной информации

Без информационные столбцы, пропущенные значения

Столбцы с идентификаторами признаков и информация об учреждениях, в которые поступили сообщения

2

Обработка временных признаков

Общий формат даты, разбиение на год, месяц, день, час

Дата / время совершения преступлений и обращений по ним

3

Обработка остальных признаков

Кодирование фичей и распределение по категориям

Возраст, пол подозреваемого / жертвы

Обработка датасета

4

Балансировка таргета

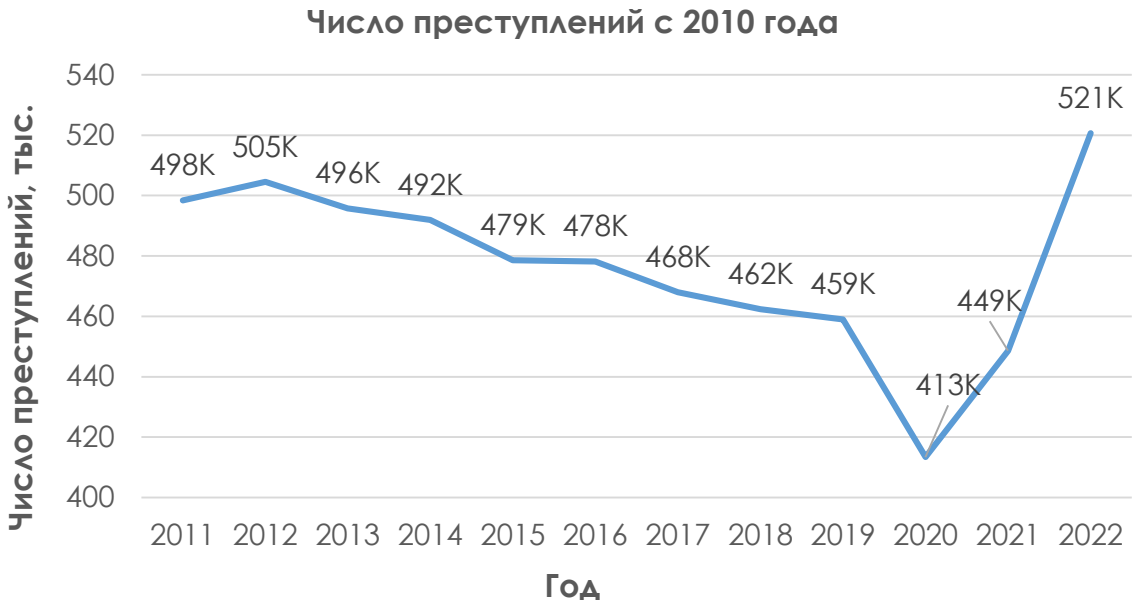
Балансировка таргета по классам преступлений

Исходный датасет содержал несблалансированные классы преступлений:

Класс	Количество
Felony (2)	4 360 957
Misdemeanor (1)	2 414 965
Violation (0)	1 035 144

Поэтому данная целевая переменная была сбалансирована специальным алгоритмом. Так, каждый класс после данной процедуры представлен в итоговом датасете в количестве 817 327 штук.

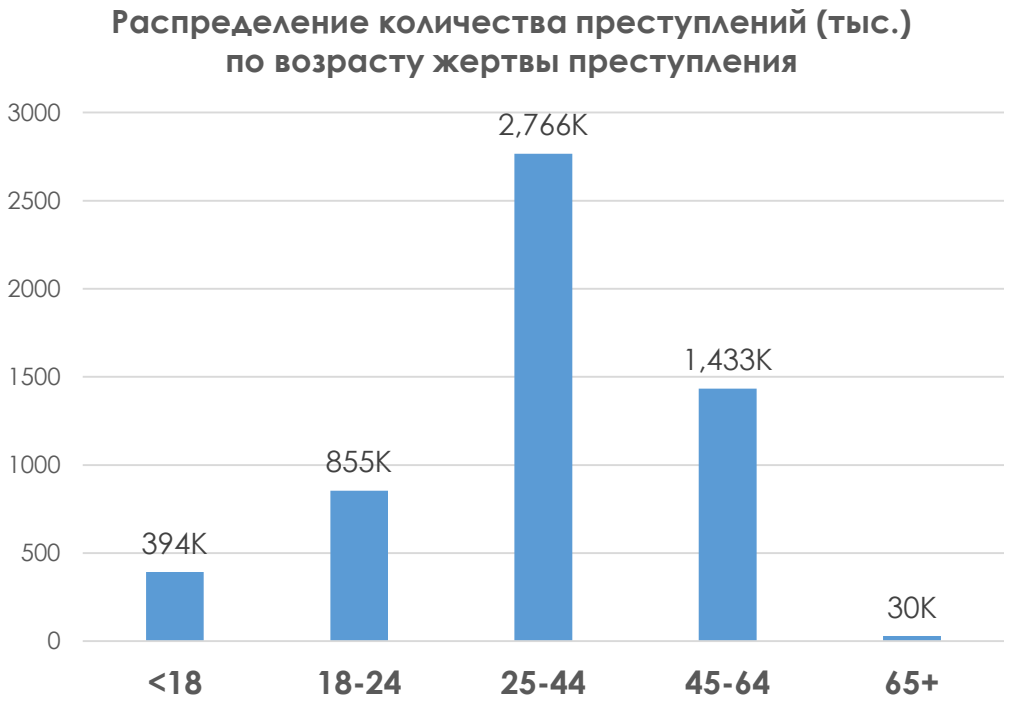
Анализ данных

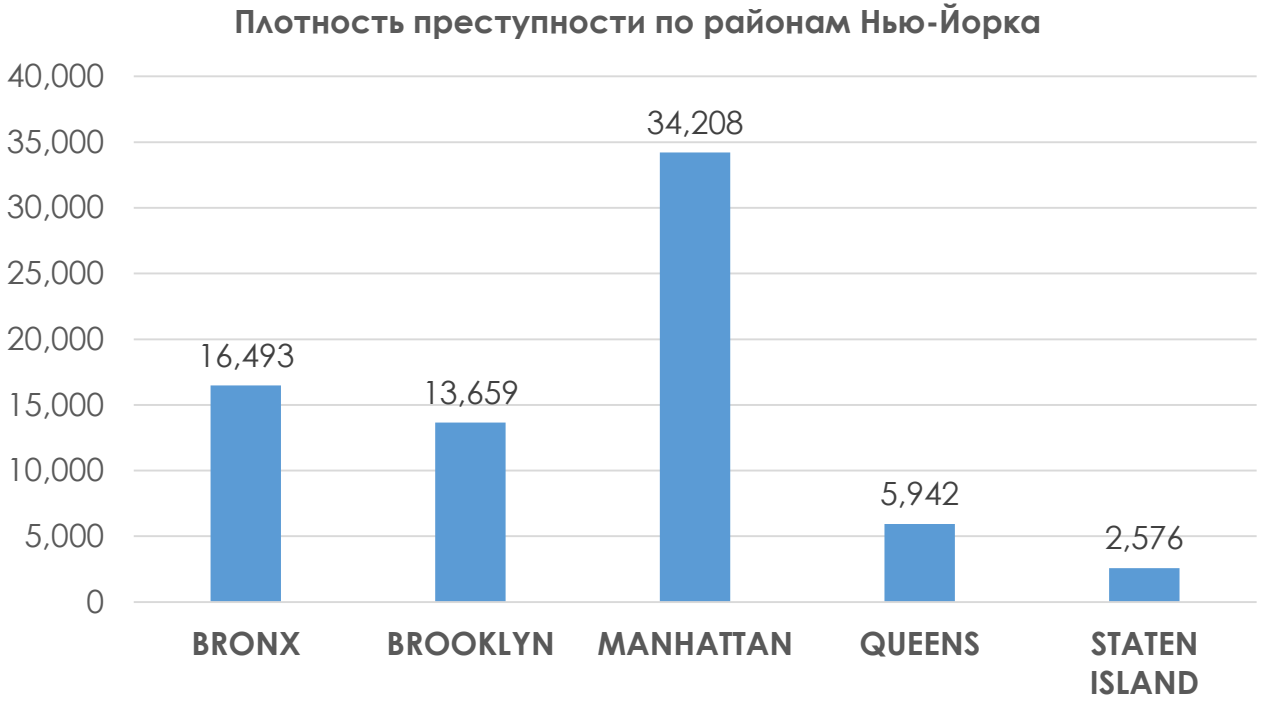


- В течение текущего десятилетия преступность шла на спад, минимум в 2020 году (вероятнее, Covid)
- С выходом из мирового карантина, преступность значительно выросла

- Наиболее часто встречаемый вид преступления – мелкая кража.
- В топ-10 входят и особо тяжкие преступления: домогательство, запрещенные вещества, нападение.

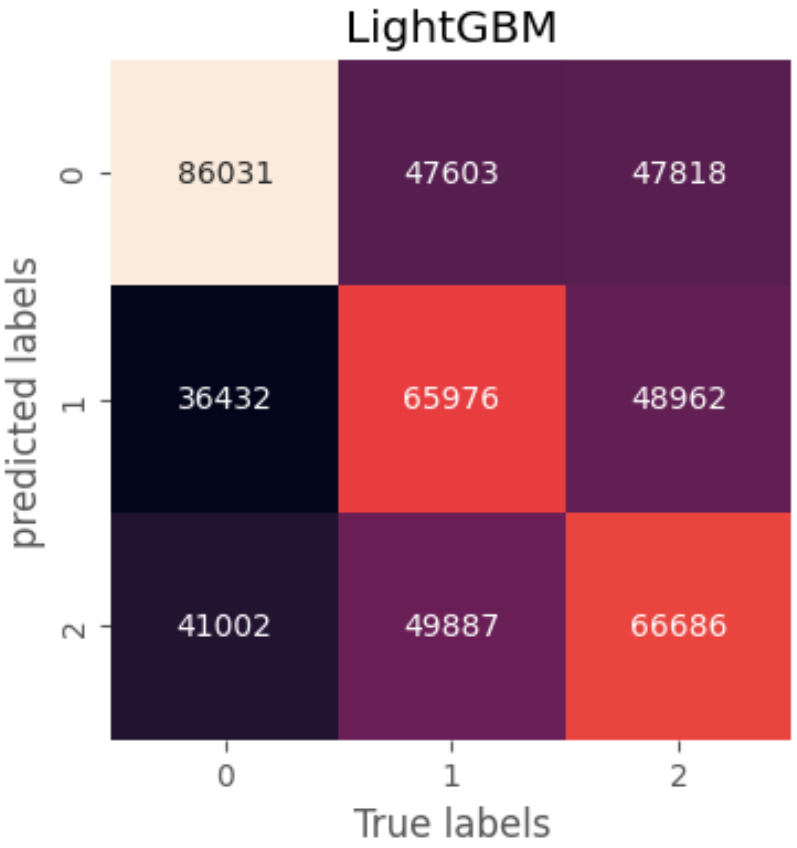
Анализ данных

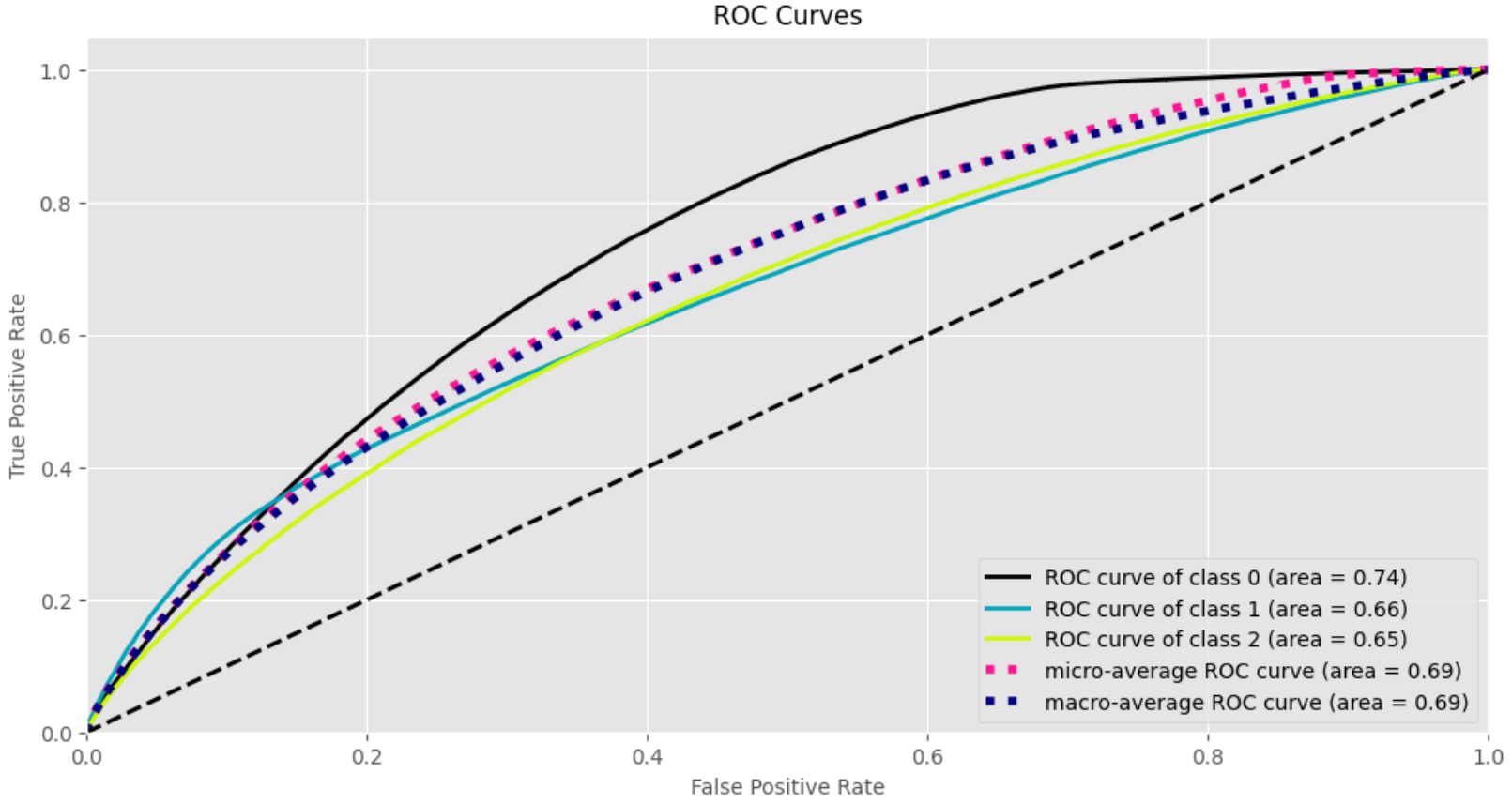




- Жертвами преступников чаще всего становятся взрослые люди от 25 до 44 лет.
- Люди более взрослого возраста менее подвержены преступным активностям.
- Наиболее преступной район по плотности – Манхэттен.
- Стейтен-айланд наименее преступной район с точки зрения плотности.

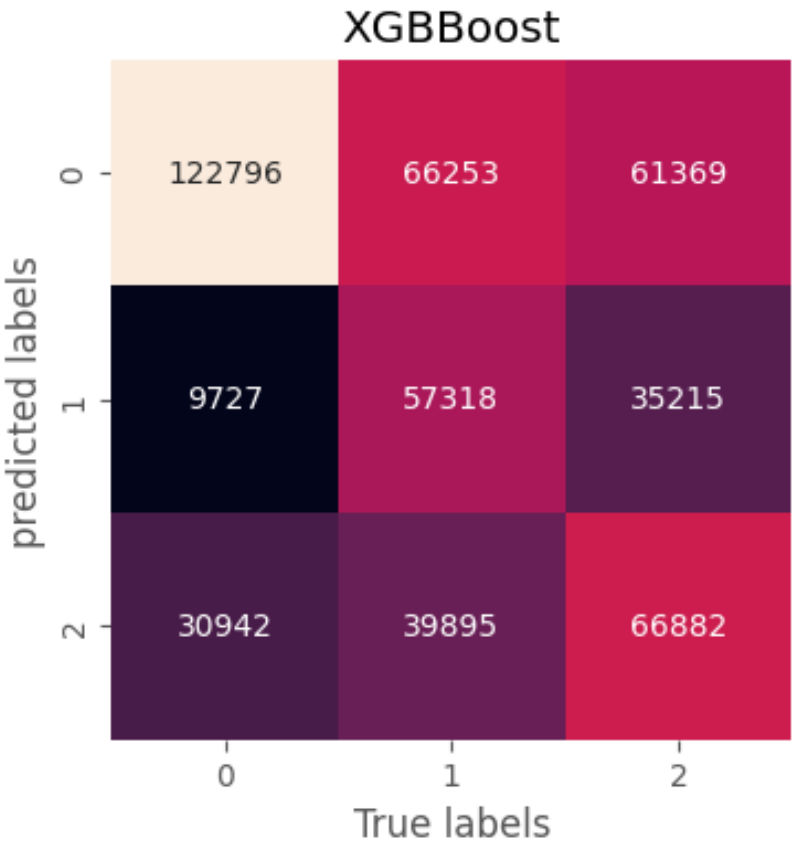
Обученные модели

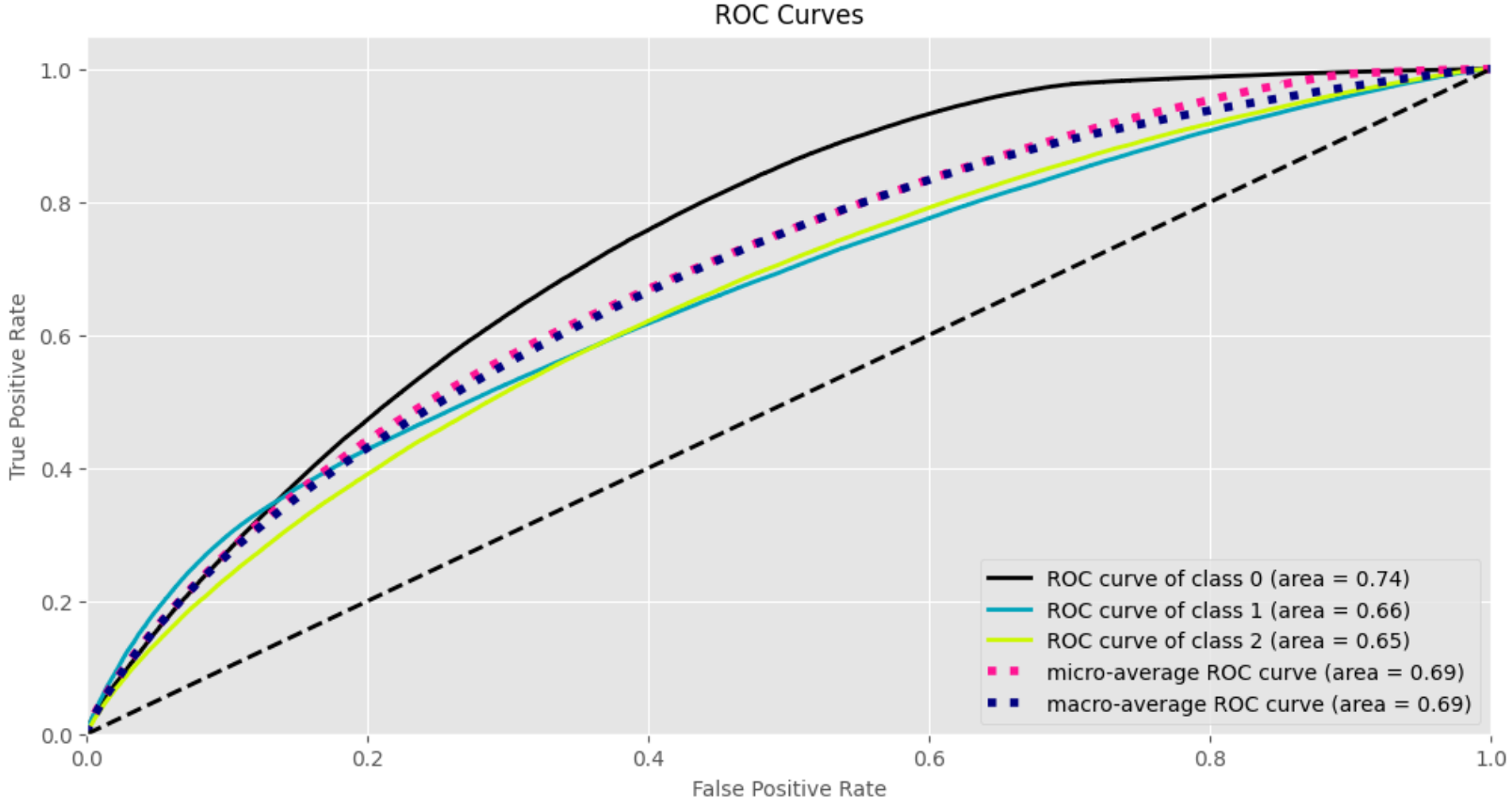




- **LightGBM Model accuracy score: 0.4460**

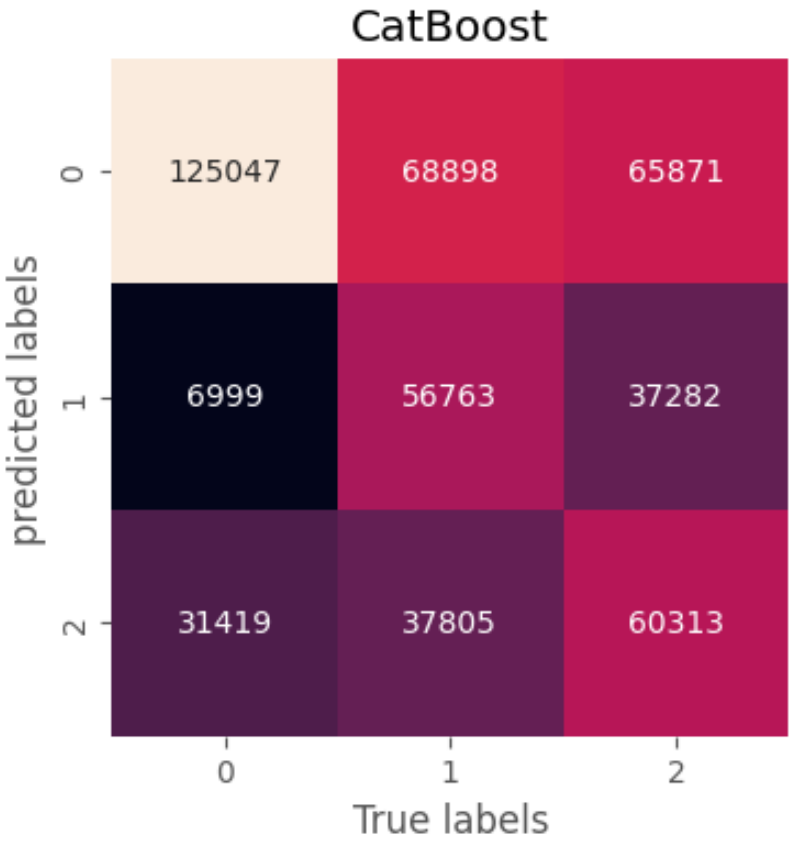
Обученные модели

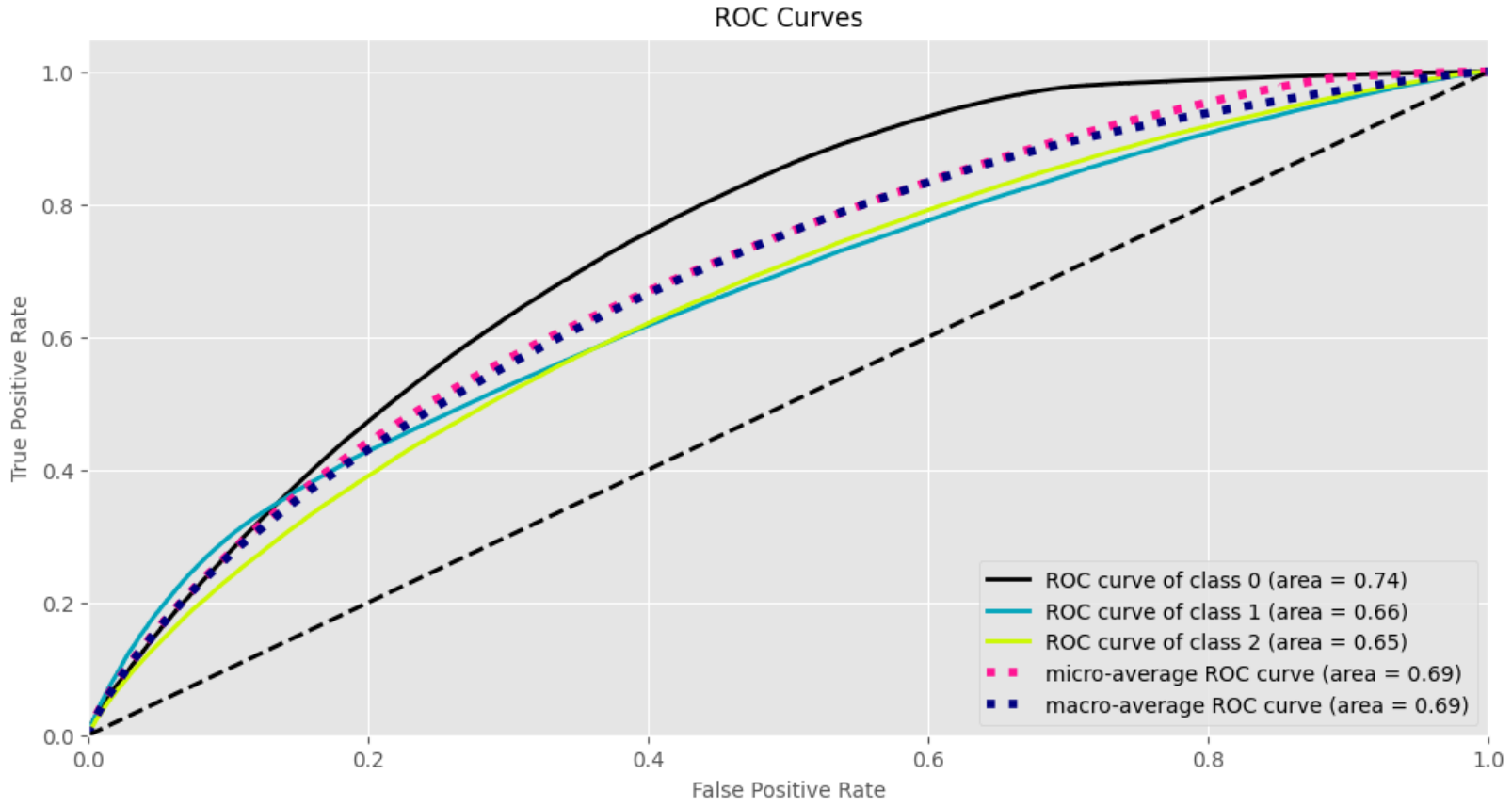




- **XGBBoost Model accuracy score: 0.5037**

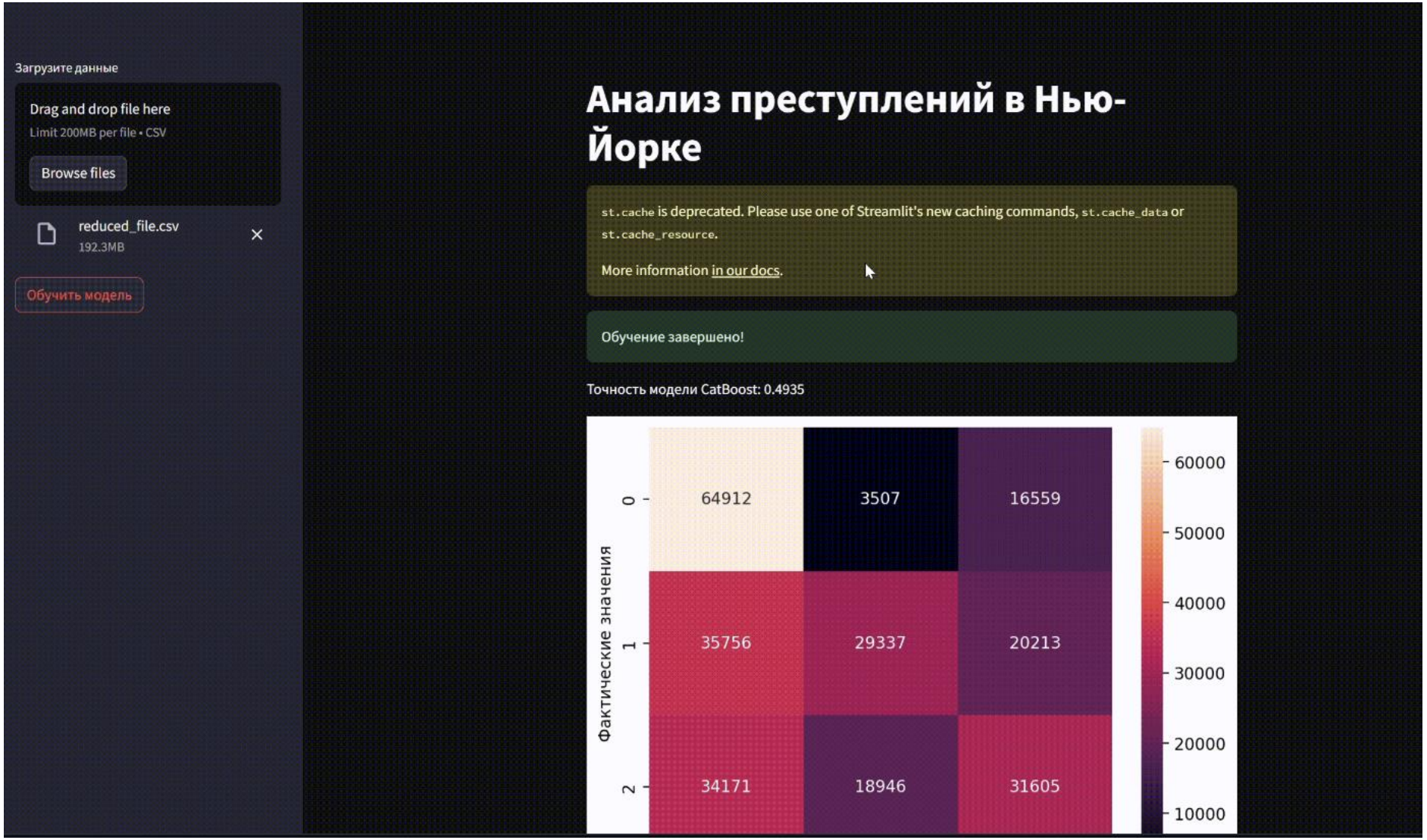
Обученные модели





- **CatBoost Model accuracy score: 0.4937**

Микросервис



Бизнес-ценность и планы

Обученная модель может помочь в развитии туризма в определенном городе / стране. С помощью описанных предсказаний туристы смогут чувствовать себя намного безопаснее и получать больше удовольствия от пребывания в незнакомом месте.

Планы для развития:

1. Расширение диапазона исходных данных: на текущий момент использованы лишь данные Нью-Йорка
2. Разработка визуального микросервиса для туристов с возможностью монетизации

Спасибо за внимание