Проект: Определение популярности геолокации для размещения банкомата

В рамках данного проекта предстояло построить модель, которая по географическим координатам и адресу точки выдаст оценку индекса популярности банкомата в этой самой локации.

Работа состоит из следующих ключевых эатпов:

1. Аналитика

2. Построение модели машинного обучения

3. Реализация сервиса

Часть 1. Аналитика

Данная часть состояла в том, чтобы изучить исходные данные о банкоматах и подготовить их к последующему обучению на них модели машинного обучения. Каждый объект-банкомат имеет следующую информацию: геоданные (широта, долгота), адрес и принадлежность к группе.

Для исходного датасета было выполнено следующее:

- проверка на пропуски;

- проверка на дубли;

- изучение описательной статистики (количественные и категориальные);

- проверка данных на валидность (например, в допустих ли пределах широта и долгота, нужной ли длины почтовый индекс):

- удаление возможных пробелов в записях.

Далее стала задача поработать с адресами расположения банкоматов: для адресов банкоматов был проведен парсинг на составляющие: почтовый индекс, страна, город, улица и дом. Для выполнения данной задачи были написаны функции, которые позволяли вычленять нужную информацию и затем использовать как значение признака о нашем объекте-банкомате.

Далее были построены графики о топ-10 городах по численности банкоматов в них и распределении нашей целевой переменной.

Завершающим в данной части было построение корреляционной матрицы числа банкоматов с городом: для этого мы применили кодировку Target Encoding для каждого города и применили сглаживание, и после рассчитали корреляцию.

Часть 2. Построение модели машинного обучения

В данной части было принято решение начать построение модели с линейных. В качестве метрик качества были выбраны R2, MAE, MAPE.

Сперва была испробована модель только на вещественных признаках, затем были приведены признаки к одному масштабу, и завершающим было добавление категориальных признаков и построение ridge (гребневой) регрессии, ее мы кроссвалидировали по 10-и фолдам и так выбрали лучшие из перебираемых гиперпараметров.

Для каждой из модели были посчитаны метрики качества и построен график для дальнейшего отслеживания улучшения нашей модели.

Часть 3. Реализация сервиса

В рамках нашего проекта была разработано приложение для анализа и предсказания популярности геолокаций, подходящих для размещения банкоматов.

Основные компоненты:

1. Веб-приложение Streamlit:

- Приложение позволяет посмотреть промежуточные результаты по выполненной работе: что было сделано, аналитику, полученные метрики качества наших моделей;

- Пользовательский интерфейс позволяет загрузить данные в формате CSV и получить предсказания;

- Приложение интегрировано с серверной частью для обработки данных и получения результатов предсказаний.

2. Серверная часть на FastAPI:

- Обрабатывает входящие данные, валидирует их, выполняет предобработку и вызывает обученную модель для получения предсказаний;

- Отправляет результат в формате CSV обратно пользователю.

Деплой приложения осуществляется командой **docker compose up -d** из корневой директории проекта. По умолчанию фронт (streamlit) будет доступен по адресу localhost:8051, бэк - по адресу localhost:8080