Определение индекса популярности банкомата

Команда:

- Вершинин Сергей
- Зайцев Антон
- Лукманова Алина

Куратор проекта:

- Гаврилова Елизавета

Постановка и данные задачи

Задача прогнозирования уровня популярности банкомата сводится к задаче **регрессии**. Значение целевой переменной является некой числовой функцией от количества операций с устройствами.

Источник данных: данные о местоположении и значении индекса популярности банкоматов с соревнования https://boosters.pro.

- **address** - адрес в транслитерации.
- **address_rus** – адрес на русском языке.
- **lat, long** - широта и долгота локации.
- **atm_group** - идентификатор банка, которому принадлежит банкомат.

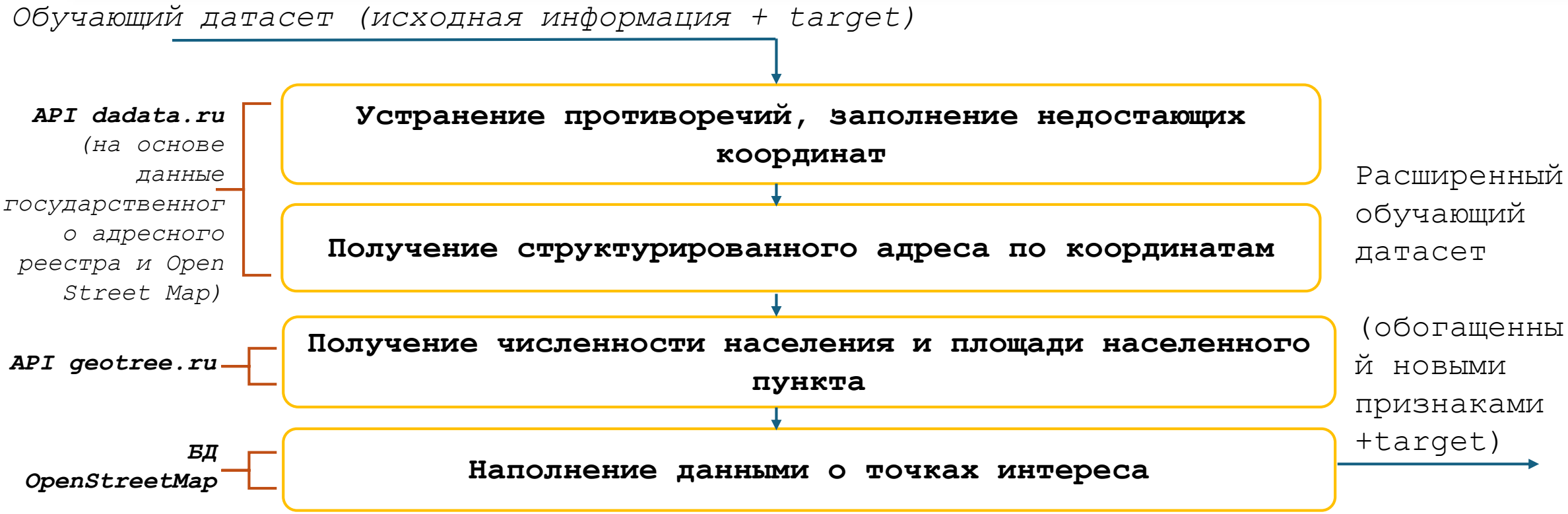
Данные задачи

Источник данных: данные с соревнования по машинному обучению, проводимого Росбанком на платформе https://boosters.pro.

Данные содержат информацию о местоположении и значении индекса популярности 6261 банкомата Росбанка и его партнеров.

- **address** - адрес в транслитерации.
- **address_rus** – адрес на русском языке.
- **lat, long** - широта и долгота локации.
- **atm_group** - идентификатор банка, которому принадлежит банкомат.

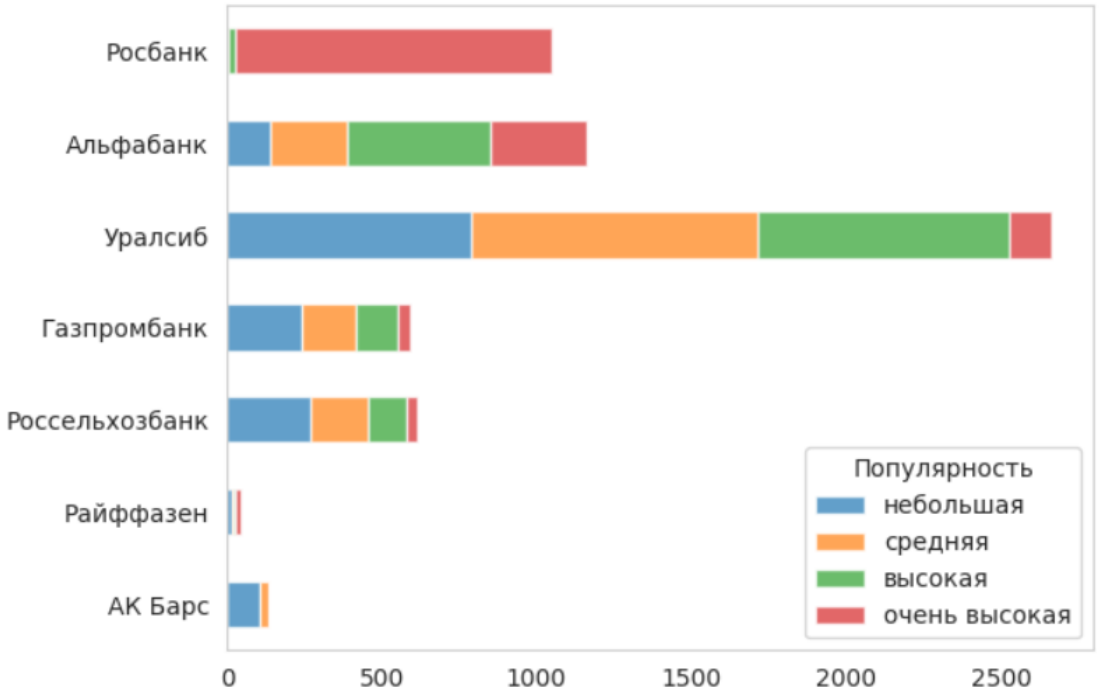
Конвейер подготовки обучающей выборки

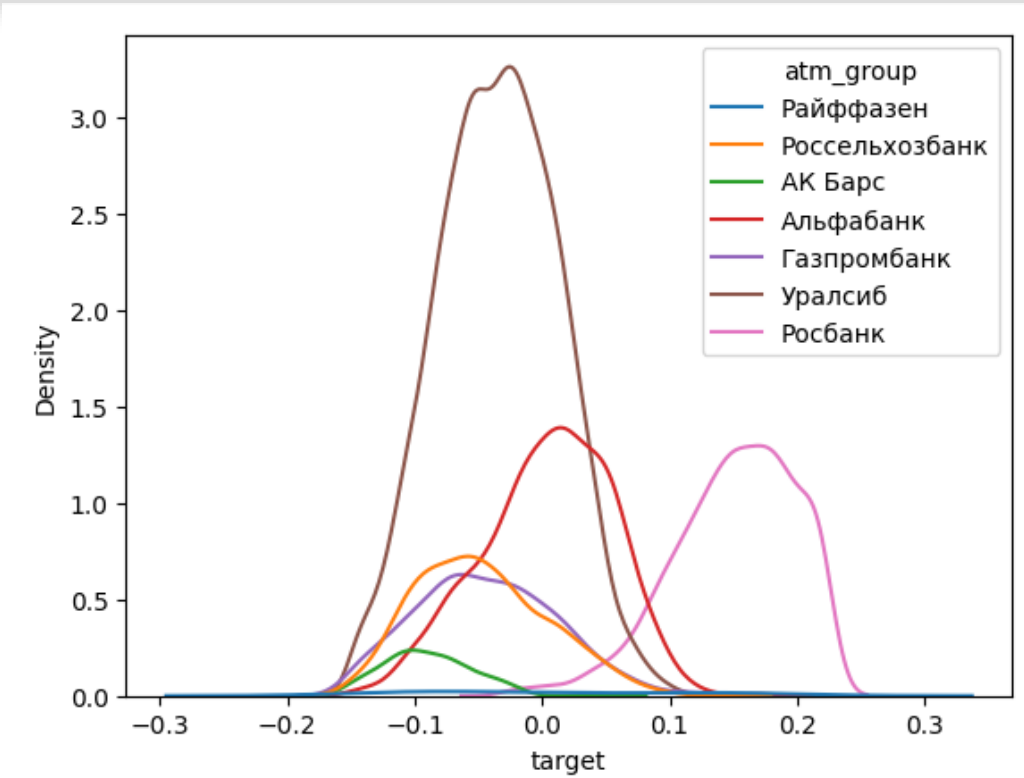


Exploratory Data Analysis

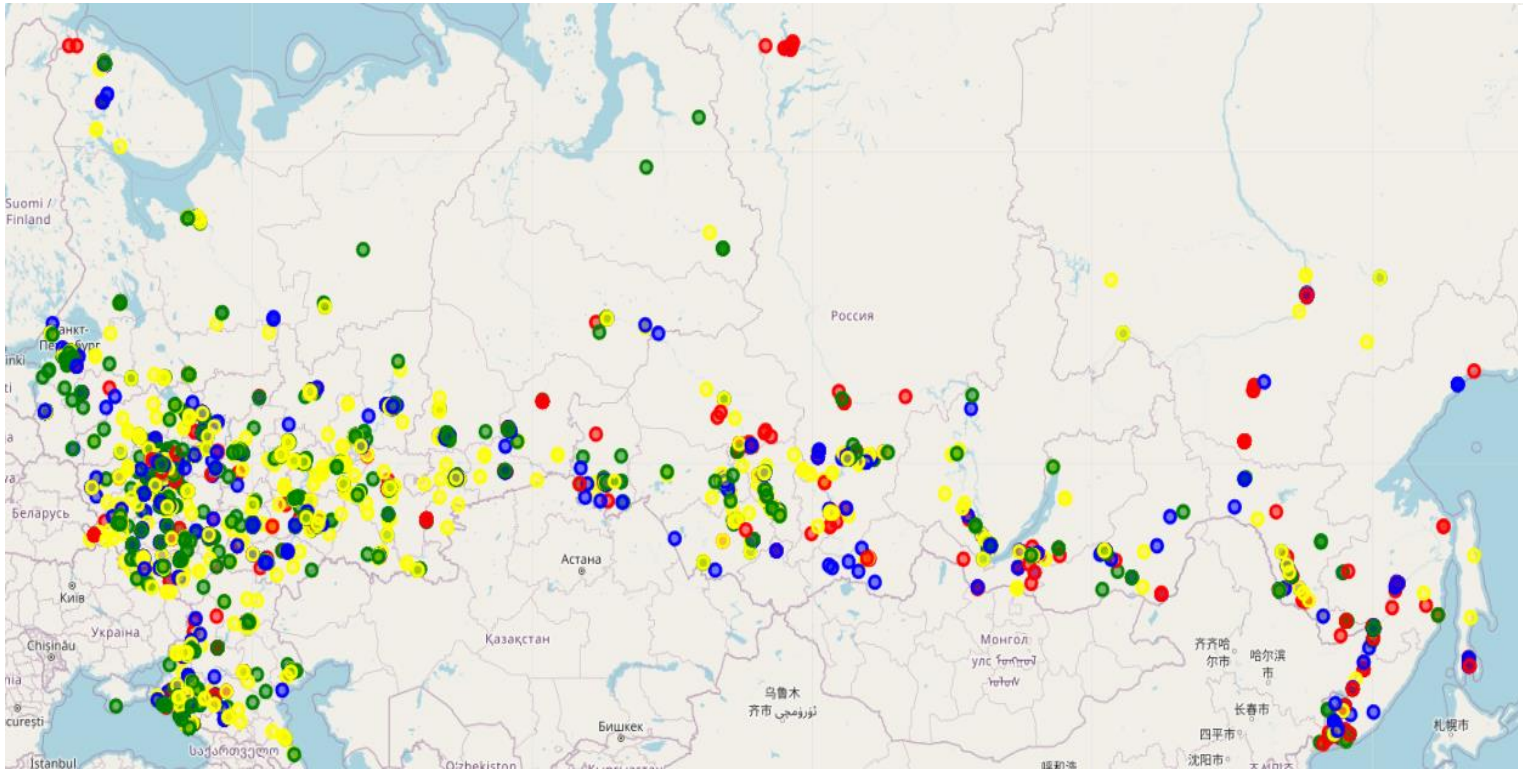
- Набор данных состоит из **6261 записей** о популярности банкоматов.
- Анализ распределения зависимости целевой переменной от категориальных признаков (банк, регион расположения).
- Анализ корреляций между числовыми признаками (выделяется связь с долготой и количеством банков поблизости)
- Информация о наличии метро поблизости было получено для 1223 банкоматов из 6261 – решено отказаться от него
- Анализ выбросов в данных, полученных на предыдущем шаге (некорректность работы сервиса DaData).

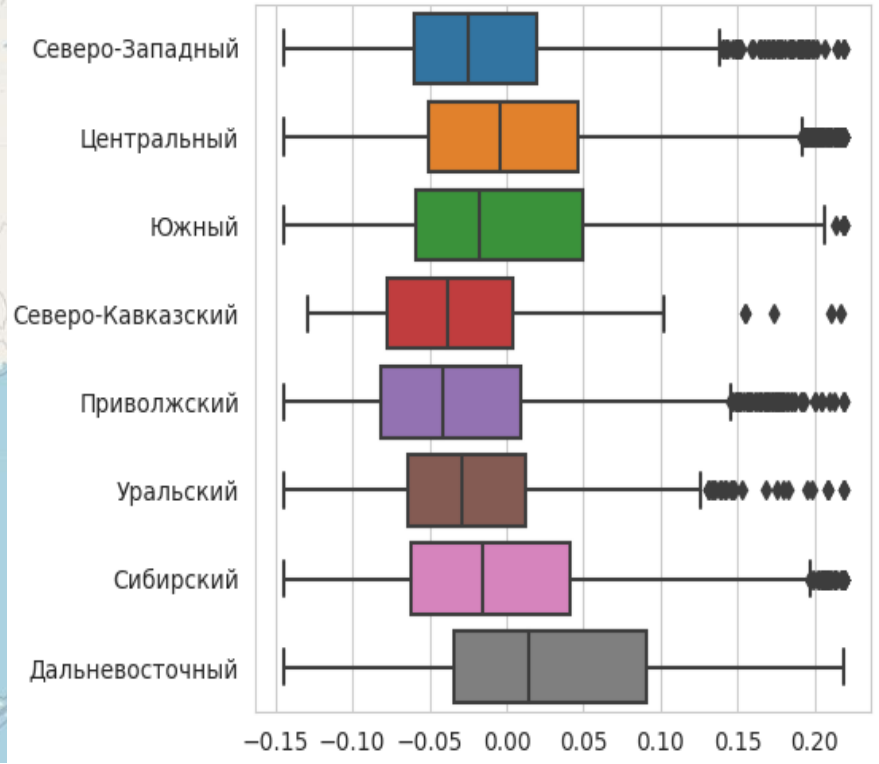
Распределение популярности банкомата у различных банков





Зависимость индекса популярности банкомата от его местоположения





0-0,25 квантиль

0,25-0,5 квантил

0,5-0,75 квантиль

0,75-1 квантиль

Чем банкомат находится восточнее или севернее, тем он более привлекателен для пользователей.

Рассмотренные модели

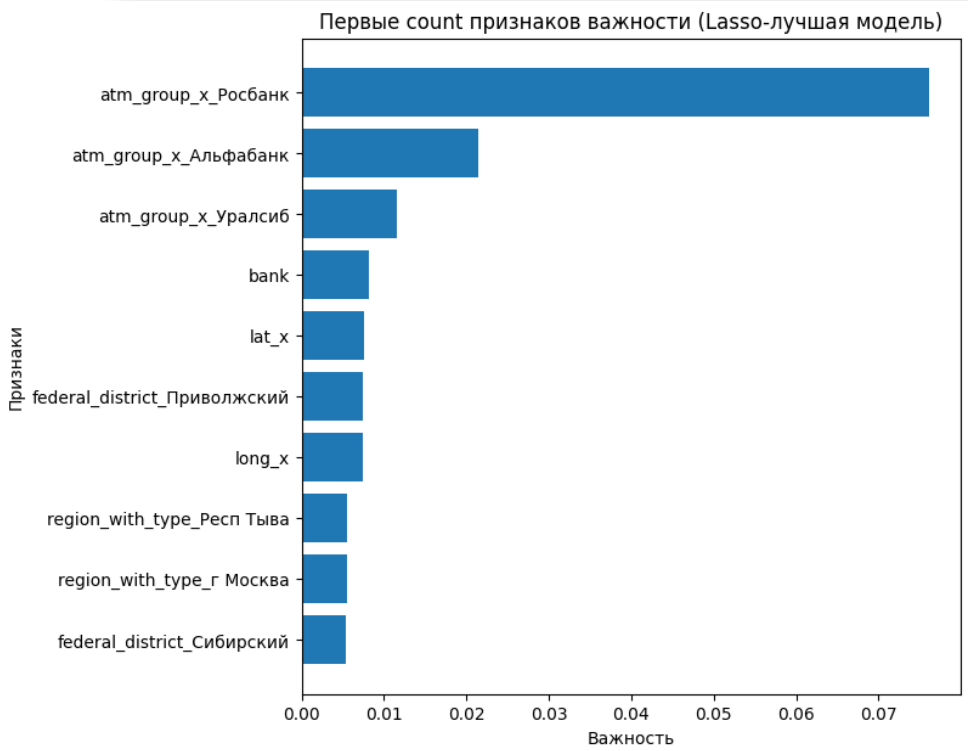
- **Линейные** модели – (Lasso, Ridge, ElasticNet) с подбором гиперпараметров.
- **Решающие** деревья – (DecisionTreeRegressor, RandomForest)
- Стекинг моделей – StackingRegressor на Lasso и Catboost.
- Модификации бустинга –(CatBoost, LGBMRegressor)

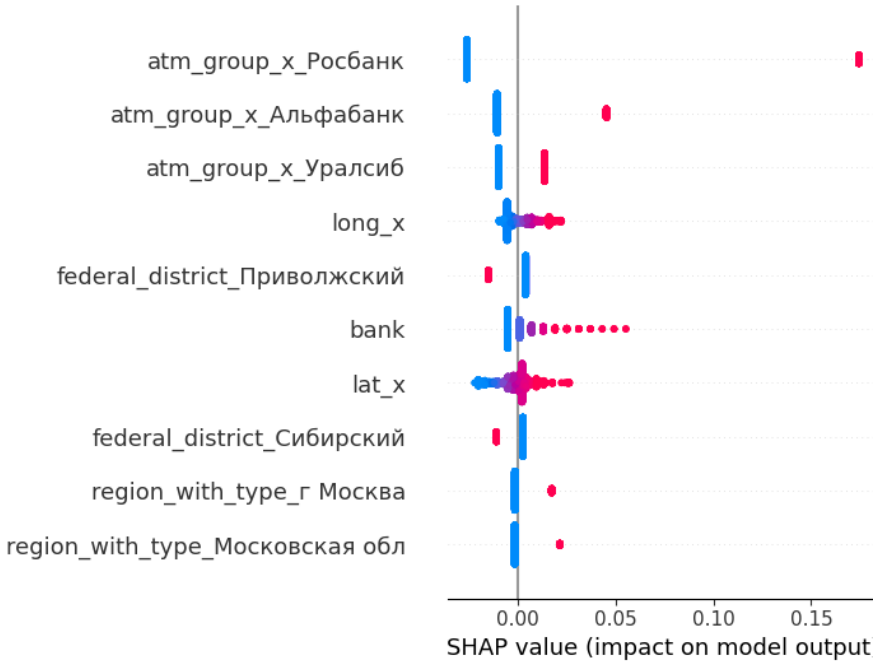
Результаты обучения моделей

	RMSE (train/test)	R2 (train/test)
Lasso (OHE)	0.0449 / 0.0469	0.7303916 0.691578
Lasso (MeanTarget)	0.0462 / 0.0472	0.714396 0.688708
RandomForest	0.0181 / 0.0442	0.955890 / 0.7261498
DecisionTreeRegressor	0.0528/ 0.0593	0.626790/ 0.508407

	RMSE (train/test)	R2 (train/test)
LGBMRegressor	0.0315 / 0.0453	0.8670599 / 0.717860
Catboost	0.0360 / 0.0445	0.826122 / 0.722630
StackingRegressor (Lasso, Catboost)	0.0311/ 0.0444	**0.870857 / 0.728667**

Анализ результатов лучшей линейной модели (Lasso, OHE)

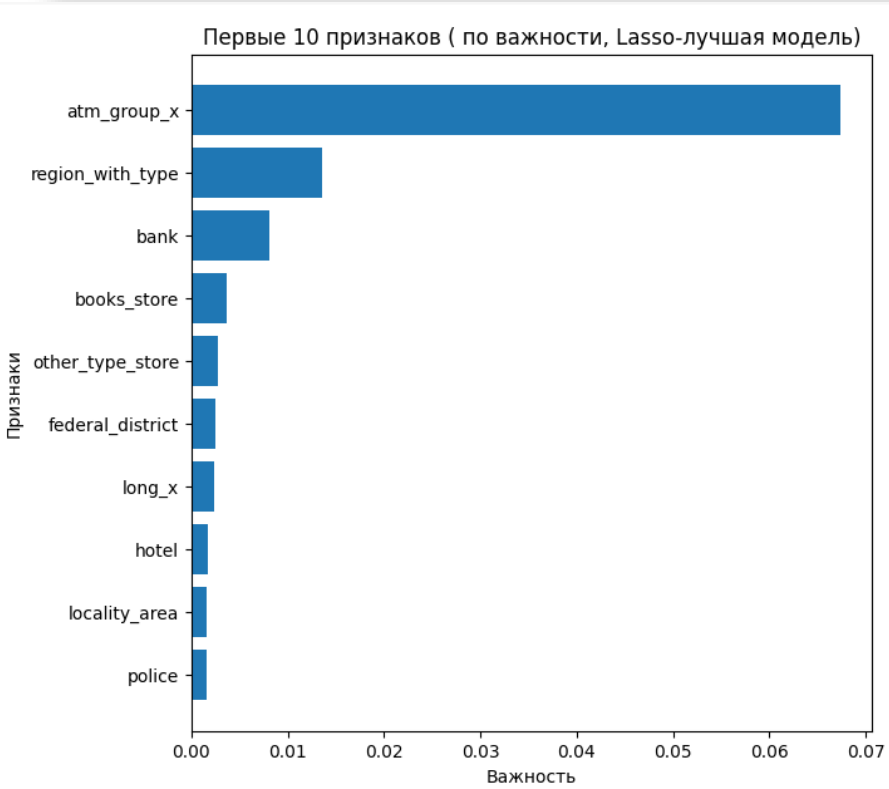


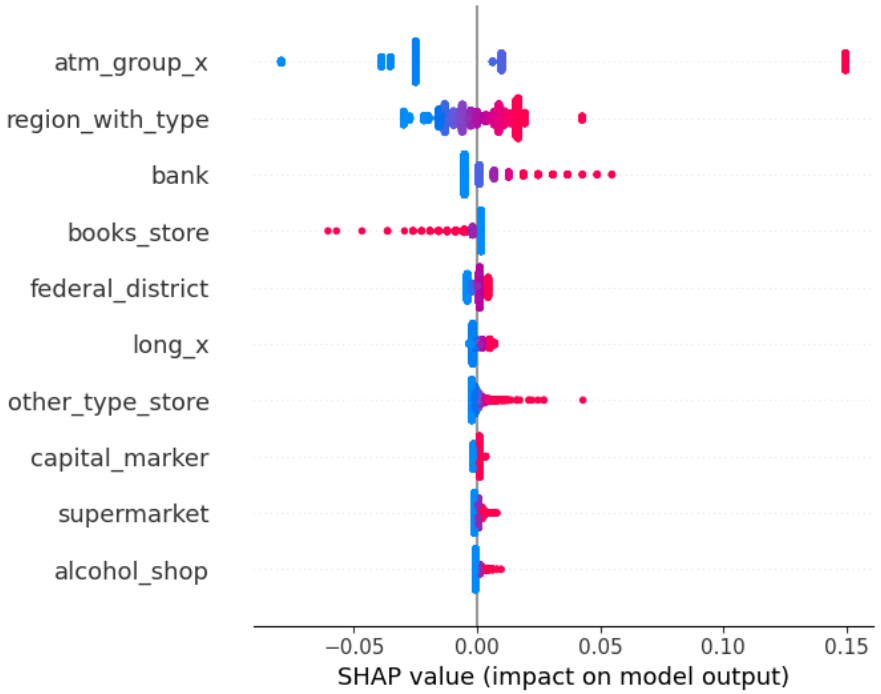


Вывод:

- определенные банки более популярны.
- Увеличение количества банков поблизости увеличивает индекс популярности.
- Чем севернее и восточнее находится банкомат, тем он популярнее.

Анализ результатов лучшей линейной модели (Lasso, MeanTarget)





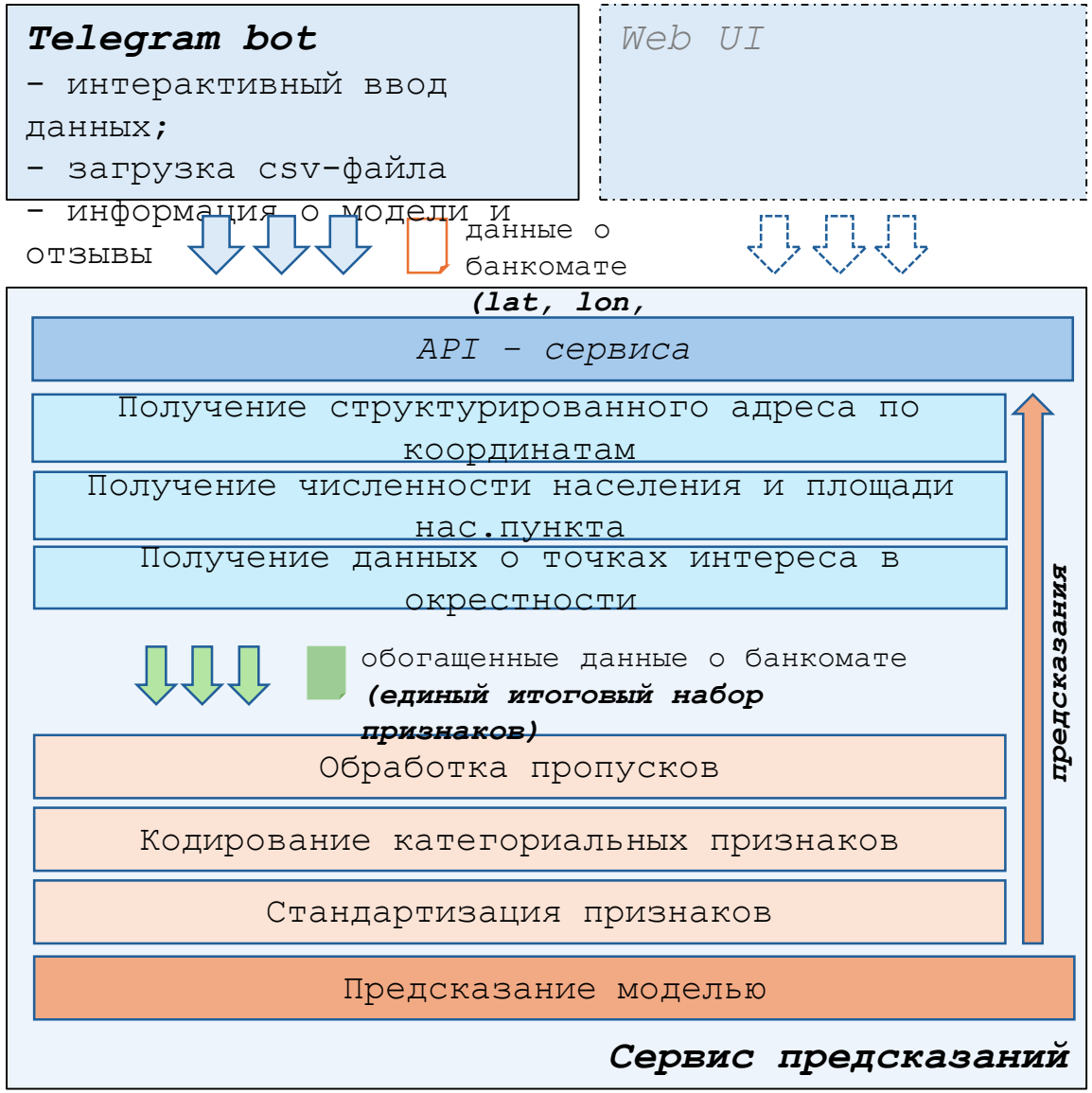
Выводы:

- Наличие супермаркетов, алкогольных и других магазинов увеличивает популярность банкомата.
- Близость Книжных магазинов – снижает.

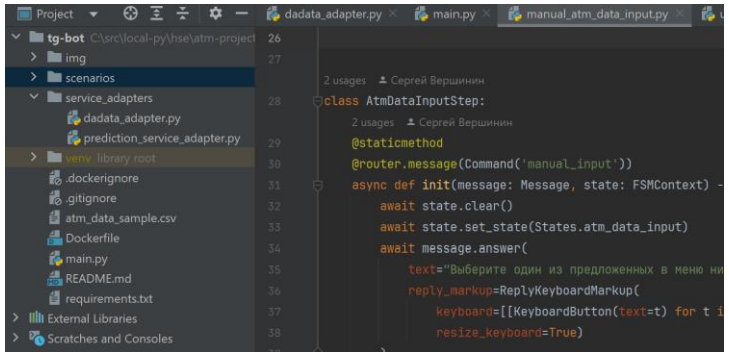
Выбор оптимальной модели

Для итоговой реализации была выбрана модель, показавшая лучшее качество - **StackingRegressor (Lasso, Catboost)**

Основные компоненты разработанного продукта



Цикл выпуска и развертывания продукта (CI/CD pipeline)



код/данные/модели на машине разработчика

unit-тестирование, проверка качества кода (github actions)

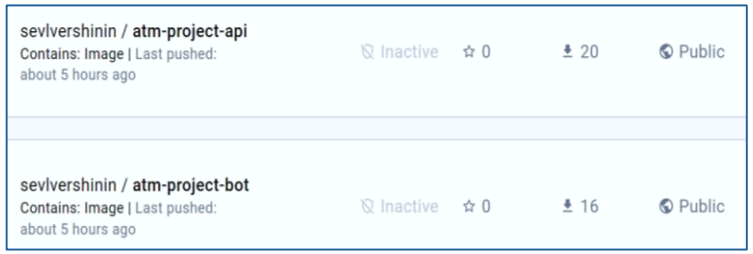
код в git-репозитории (*https://github.com/.../atm-project*)

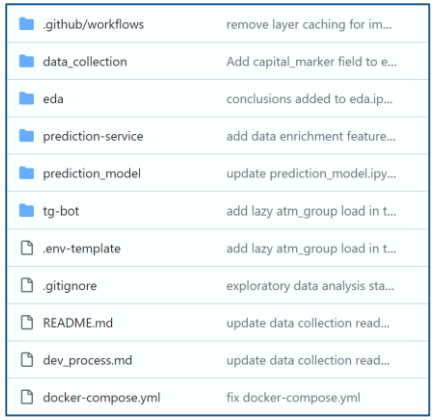
сборка docker-образов, размещение в Docker Hub (github actions, release published trigger)

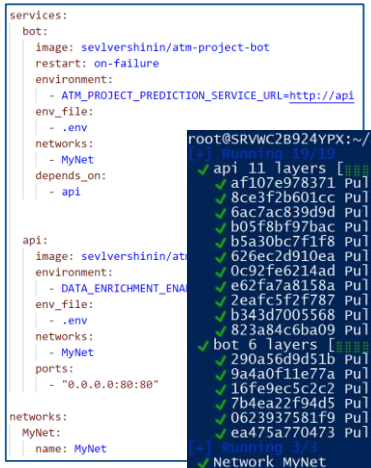
образы компонент на Docker Hub
https://hub.docker.com/.../atm-project-api/
https://hub.docker.com/.../atm-project-bot/

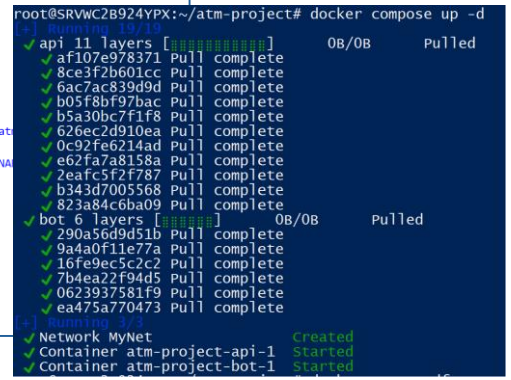
загрузка/обновление образов, запуск системы (docker compose pull, docker

контейнеры с компонентами системы на ВМ
OS Ubuntu









Планы развития

На следующих этапах развития проекта планируется:

- разработать Web-клиент для более удобного взаимодействия пользователей с сервисом;
- выделить функциональность по обогащению входных данных пользователя в отдельный сервис;
- настроить механизмы мониторинга работающего сервиса (логирование, метрики и т.п.);
- внедрить в процесс разработки практики MLOps;
- посмотреть другие ансамблевые модели, попытаться подобрать гиперпараметры.