

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

Обучающий датасет (исходная информация + target)

API *dadata.ru*

(на основе
данные

государственног
о адресного
реестра и Open
Street Map)

Устранение противоречий, заполнение недостающих
координат

Получение структурированного адреса по координатам

API *geotree.ru*

Получение численности населения и площади населенного
пункта

БД
OpenStreetMap

Наполнение данными о точках интереса

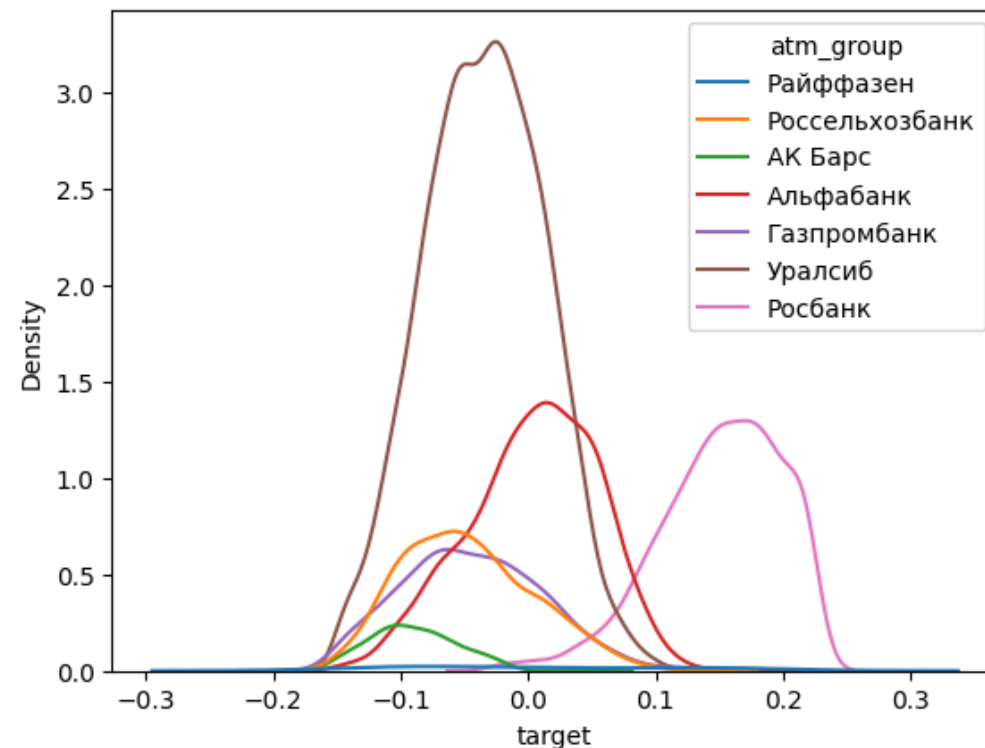
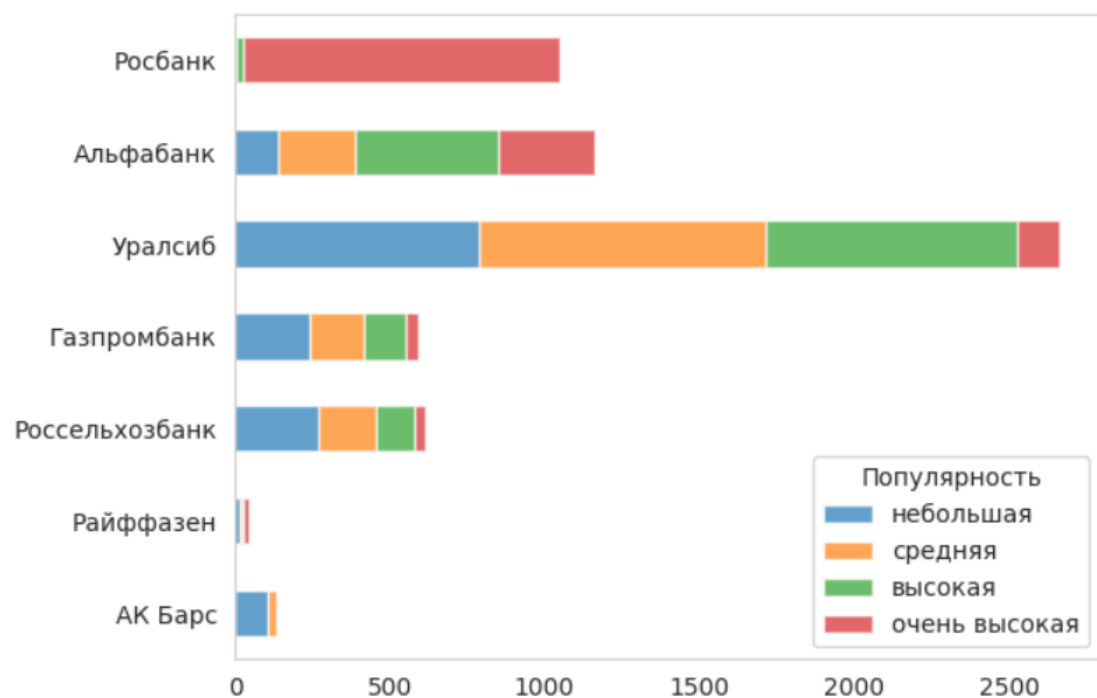
Расширенный
обучающий
датасет

(обогащенны
й новыми
признаками
+target)

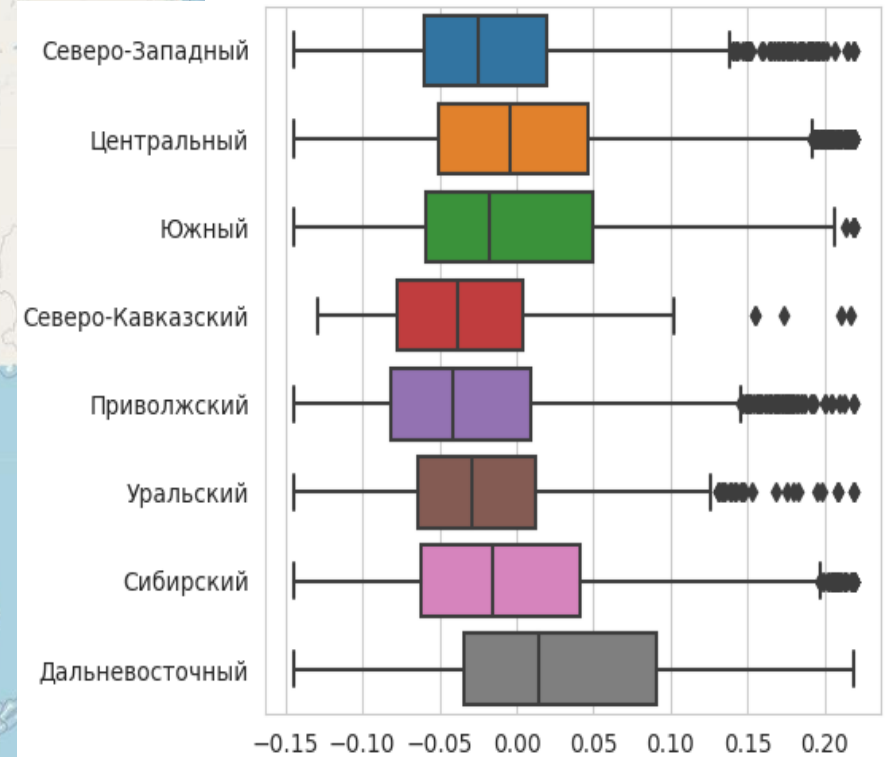
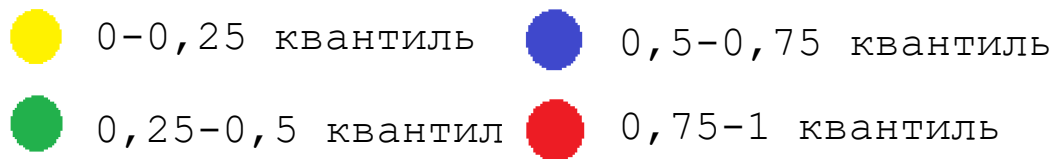
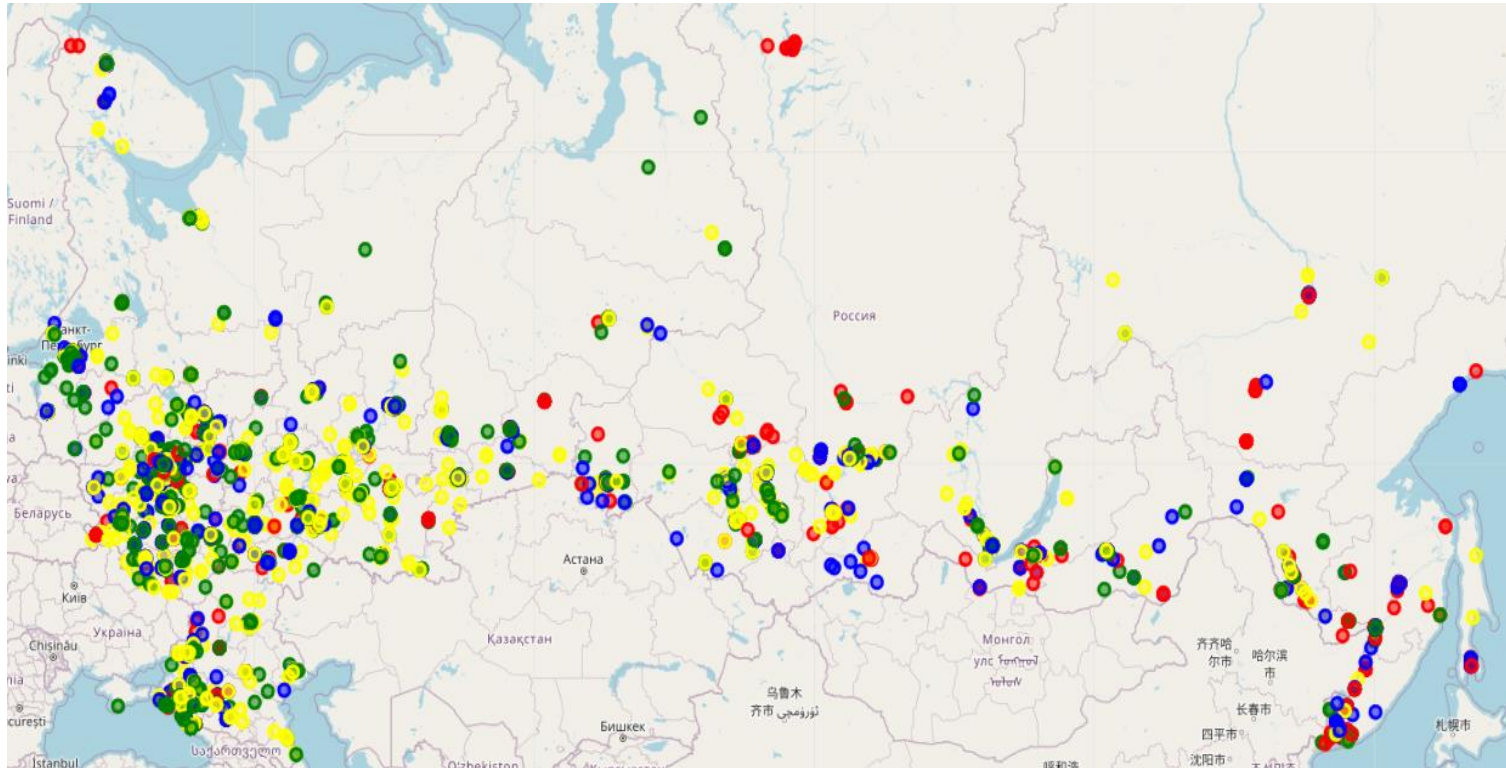
Exploratory Data Analysis

- 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



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



Рассмотренные модели

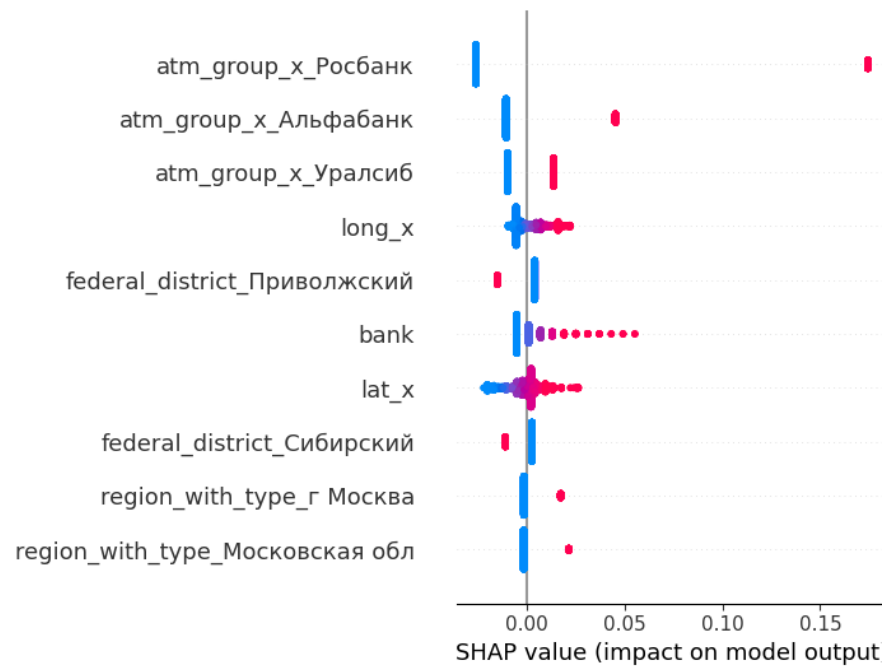
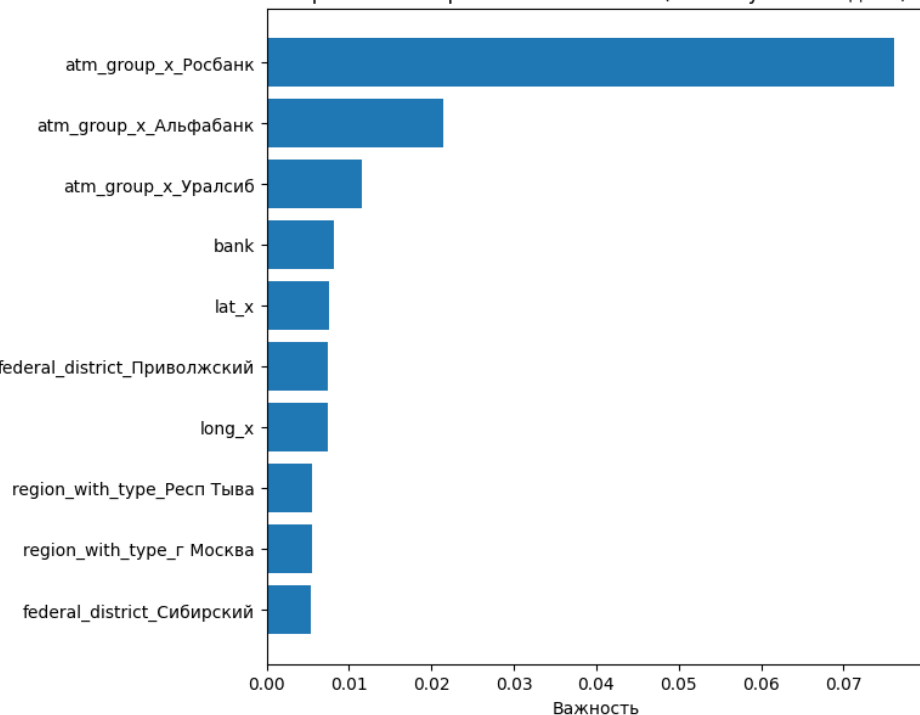
- **Линейные** модели – (Lasso, Ridge, ElasticNet) с подбором гиперпараметров.
 - **Решающие** деревья – (DecisionTreeRegressor, RandomForest)
 - Стекинг моделей – StackingRegressor на Lasso и Catboost.
 - Модификации бустинга – (CatBoost, LGBMRegressor)
-

Результаты обучения моделей

	RMSE (train/test)	R2 (train/test)		RMSE (train/test)	R2 (train/test)
Lasso (OHE)	0.0449 / 0.0469	0.7303916 0.691578	LGBMRegressor	0.0315 / 0.0453	0.8670599 / 0.717860
Lasso (MeanTarget)	0.0462 / 0.0472	0.714396 0.688708	Catboost	0.0360 / 0.0445	0.826122 / 0.722630
RandomForest	0.0181 / 0.0442	0.955890 / 0.7261498	StackingRegressor (Lasso, Catboost)	0.0311/ 0.0444	0.870857 / 0.728667
DecisionTreeRegressor	0.0528/ 0.0593	0.626790/ 0.508407			

Анализ результатов лучшей линейной модели (Lasso, ONE)

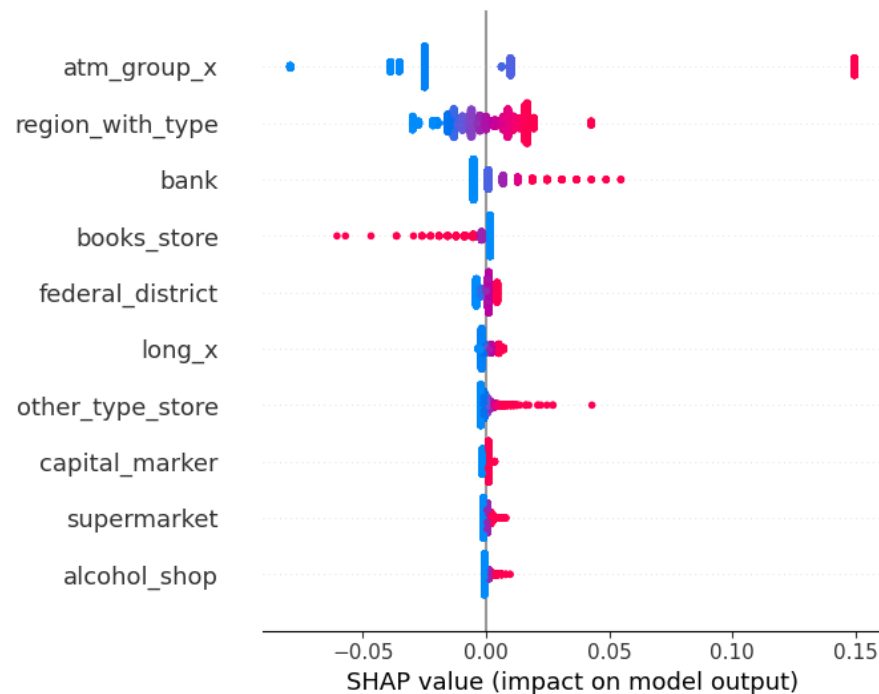
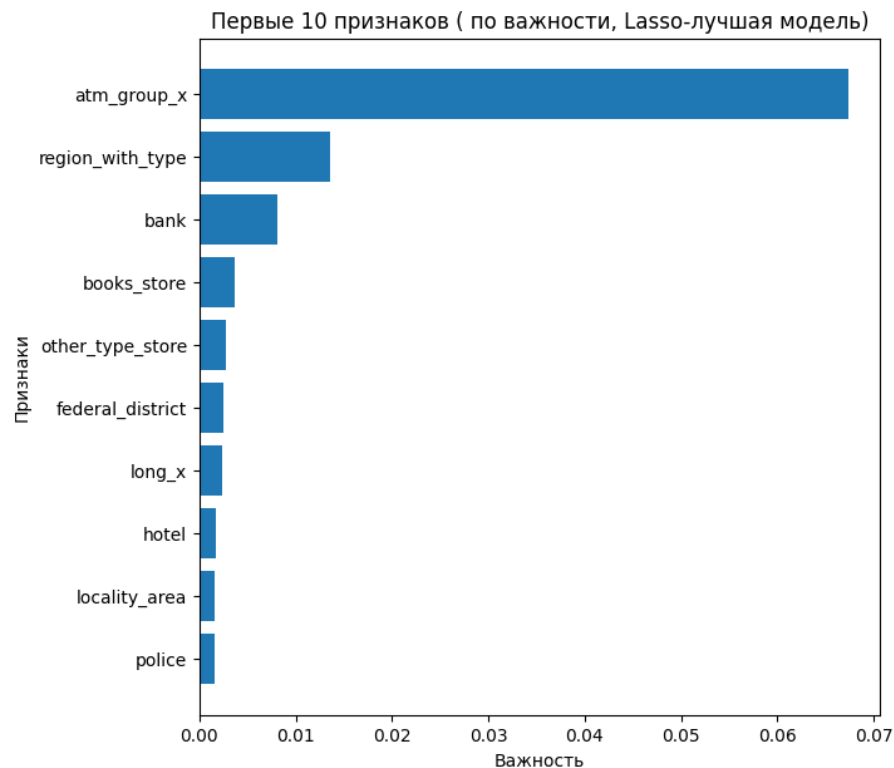
Первые count признаков важности (Lasso-лучшая модель)



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:

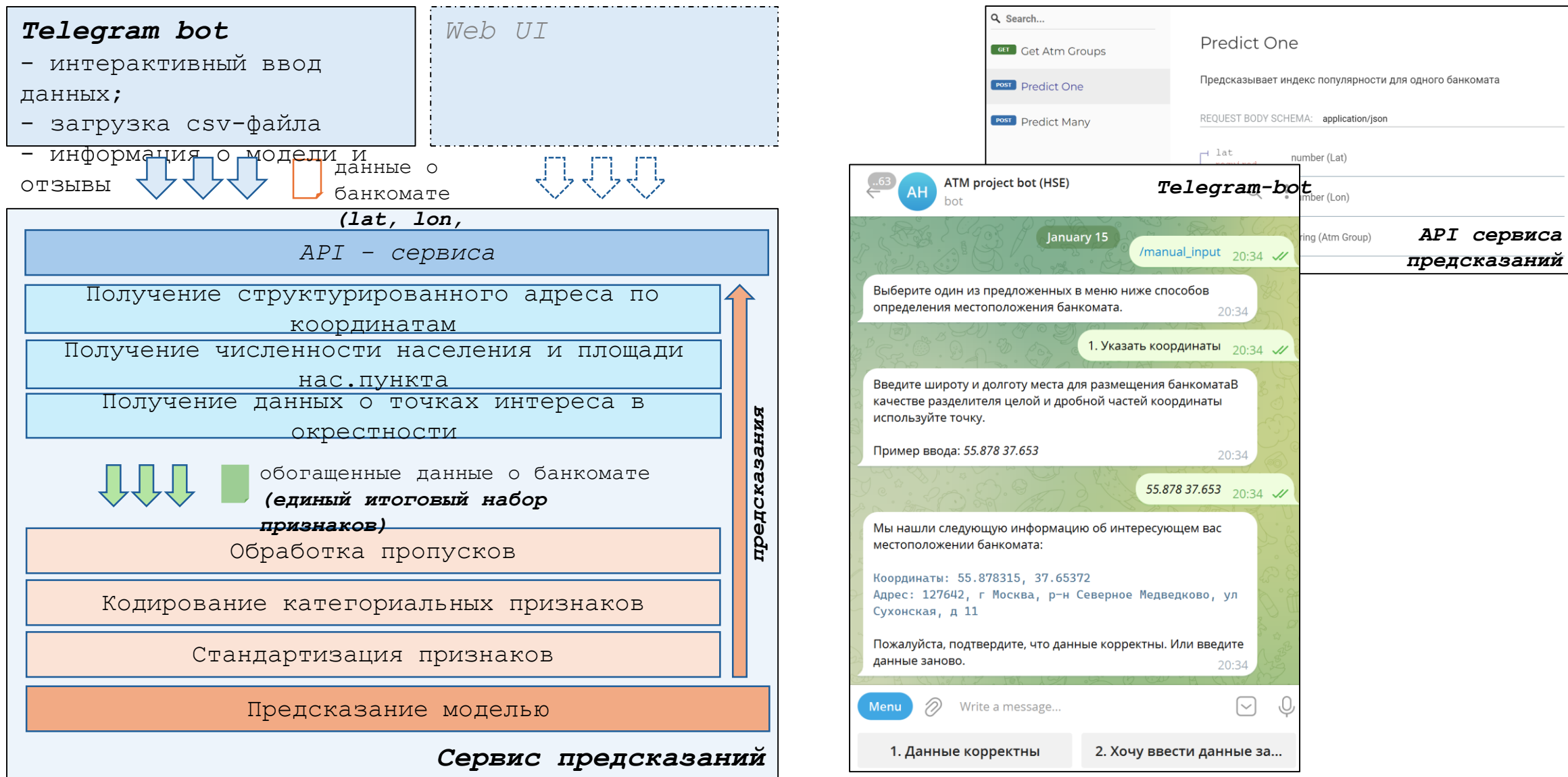
- Наличие супермаркетов, алкогольных и других магазинов увеличивает популярность банкомата.
- Близость Книжных магазинов – снижает.



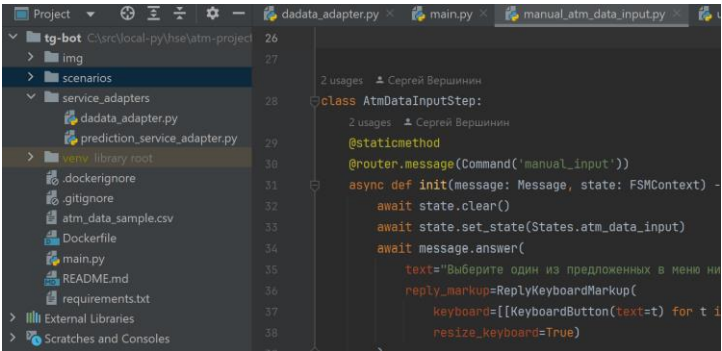
Выбор оптимальной модели

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 VM OS Ubuntu

.github/workflows	remove layer caching for im...
data_collection	Add capital_marker field to e...
eda	conclusions added to eda.ip...
prediction-service	add data enrichment feature...
prediction_model	update prediction_model.py...
tg-bot	add lazy atm_group load in t...
.env-template	add lazy atm_group load in t...
.gitignore	exploratory data analysis sta...
README.md	update data collection read...
dev_process.md	update data collection read...
docker-compose.yml	fix docker-compose.yml

sevlvershinin / atm-project-api	Inactive	0	20	Public
Contains: Image Last pushed: about 5 hours ago				
sevlvershinin / atm-project-bot	Inactive	0	16	Public
Contains: Image Last pushed: about 5 hours ago				

```
services:
  bot:
    image: sevlvershinin/atm-project-bot
    restart: on-failure
    environment:
      - ATM_PROJECT_PREDICTION_SERVICE_URL=http://api
    env_file:
      - .env
    networks:
      - MyNet
    depends_on:
      - api
  api:
    image: sevlvershinin/atm-project-api
    environment:
      - DATA_ENRICHMENT_ENA
    env_file:
      - .env
    networks:
      - MyNet
    ports:
      - "0.0.0.0:80:80"
networks:
  MyNet:
    name: MyNet
```

```
root@SRVWC2B924YPX:~/atm-project# docker compose up -d
Running 13/10
api 11 layers [#####] 0B/0B Pulled
  af107e978371 Pull complete
  8ce3f2b601cc Pull complete
  8ac7ac839d9d Pull complete
  b05f8bf97bac Pull complete
  b5a30bc7f1f8 Pull complete
  626ec2d910ea Pull complete
  0c92fe6214ad Pull complete
  e62fa7a8158a Pull complete
  2eafc5f2f787 Pull complete
  b343d7005568 Pull complete
  823a84c6ba09 Pull complete
bot 6 layers [#####] 0B/0B Pulled
  290a56d9d51b Pull complete
  9a4a0f11e77a Pull complete
  16fe9e5e2c22 Pull complete
  7b4ea22f94d5 Pull complete
  0623937581f9 Pull complete
  ea475a770473 Pull complete
Network MyNet Created
Container atm-project-api-1 Started
Container atm-project-bot-1 Started
```

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.
-