ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)» (МАИ)

Project for the discipline «ML System Design»

«Apartment_rent_price»

Developed by:
Брежнев В. А.
Пешков М.Ю.

Problem

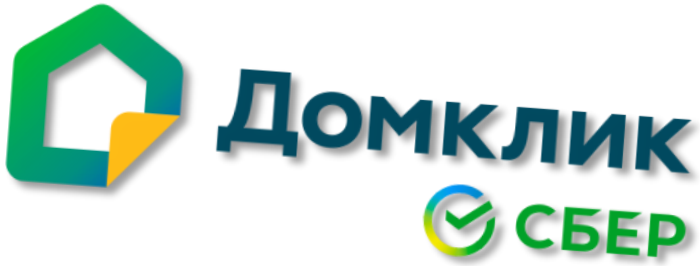
The real estate market is one of the key markets in the global economy.

The rise in property prices, as well as rental costs, is one of the major challenges, especially given the slow growth of population wages.

People strive to find the most rational way to maximize benefits when searching for housing, spending minimal funds while still obtaining acceptable comfort conditions.

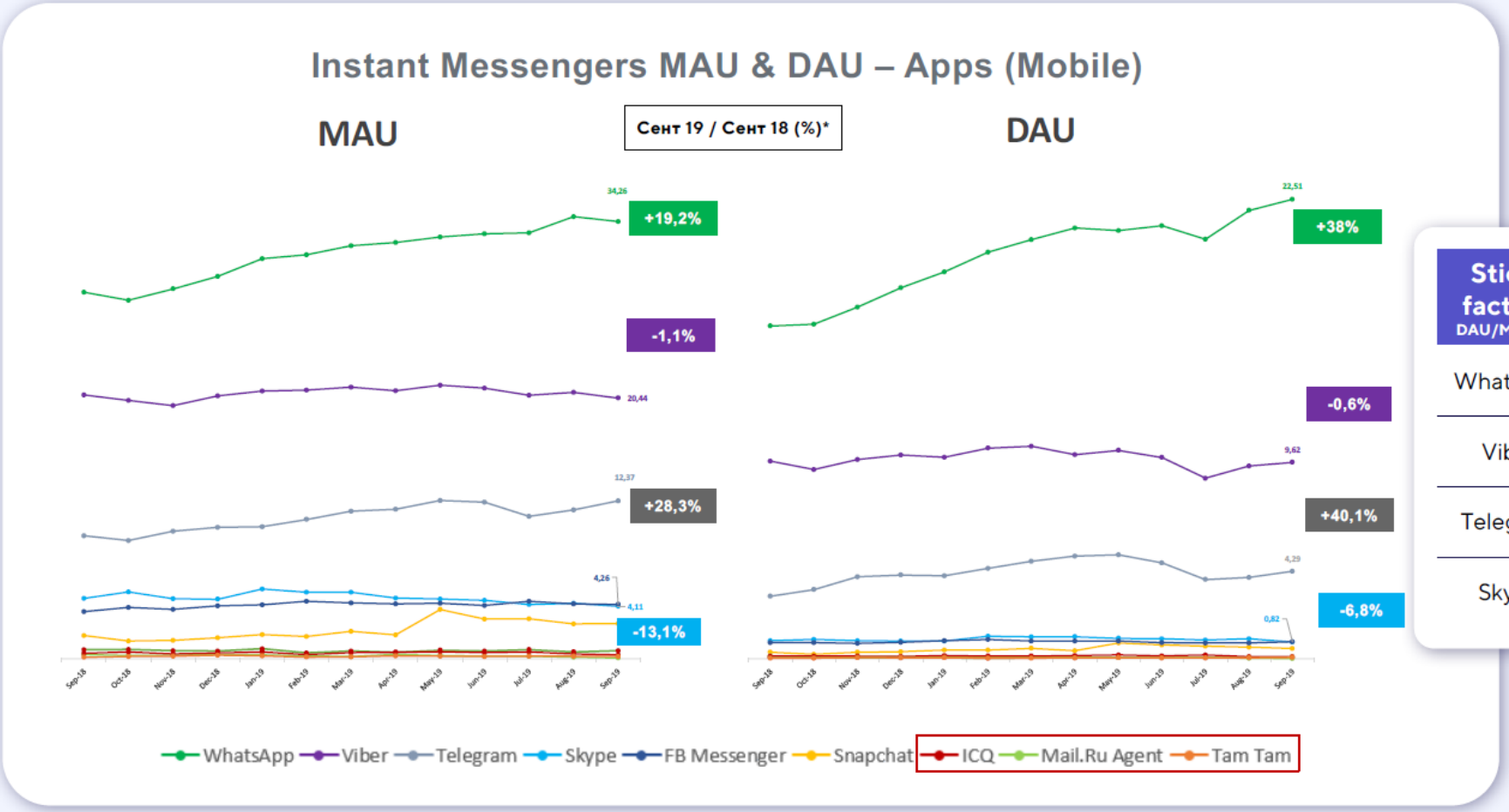
Numerous factors play a decisive role in shaping rental prices. Applying various analytical methods can help better understand the key factors influencing rental costs, as well as assist in finding the most rational offer available on the real estate market.

Market



Messengers

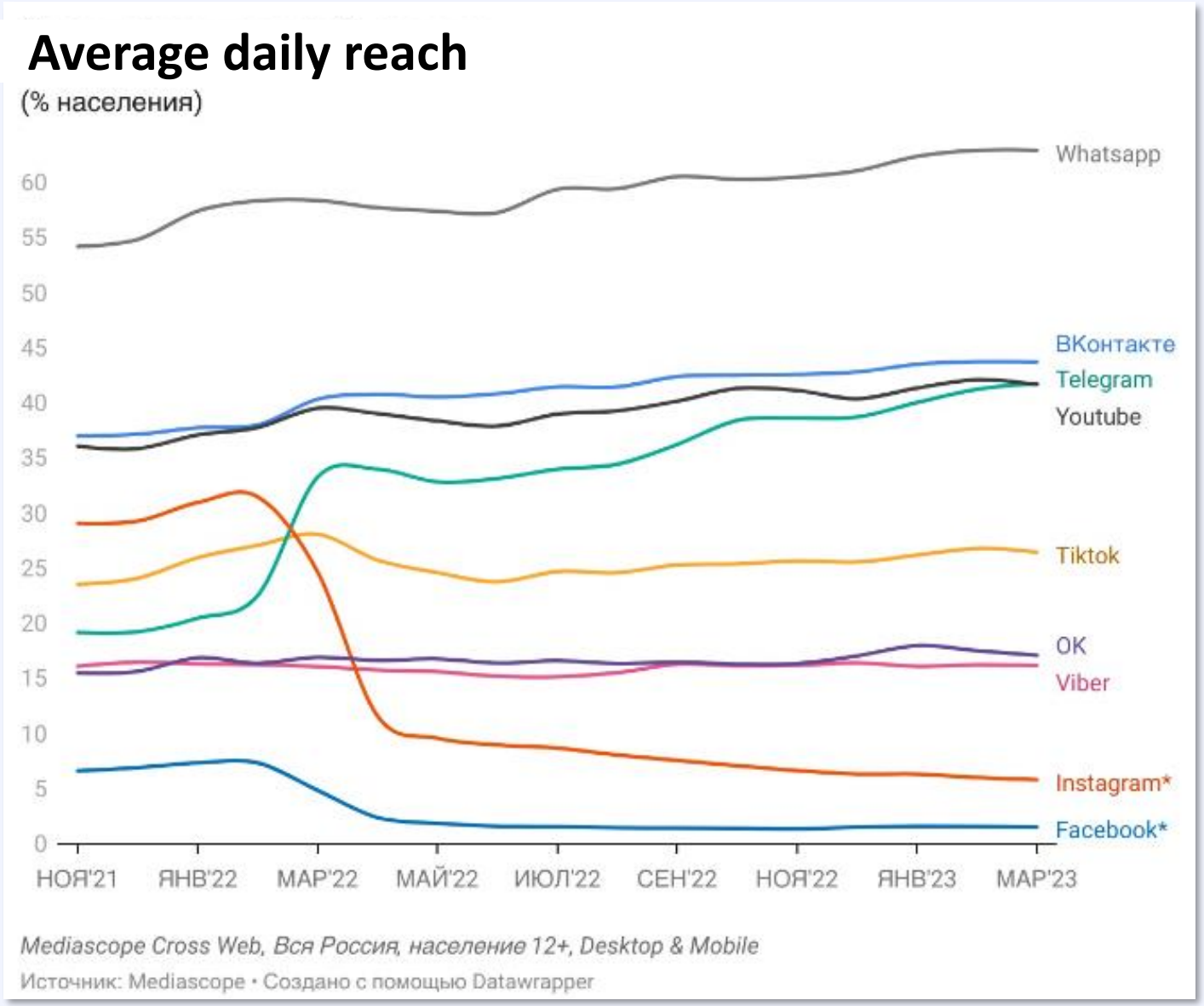
Messaging apps demonstrate strong year-over-year growth



Sticky factor = DAU/MAU (%)	Sept19
WhatsApp	65,7%
Viber	47%
Telegram	34,6%
Skype	19,9%

Источник: Mediascope – Russia, cities 100k+, 12-64, unique users, mln

Messengers


Average daily reach
(% населения)
60
55
50
45
40
35
30
25
20
15
10
5
0
НОЯ'21
ЯНВ'22
МАР'22
МАЙ'22
ИЮЛ'22
СЕН'22
НОЯ'22
ЯНВ'23
МАР'23
Whatsapp
ВКонтакте
Telegram
Youtube
Tiktok
OK
Viber
Instagram*
Facebook*
Mediascope Cross Web, Вся Россия, население 12+, Desktop & Mobile
Источник: Mediascope • Создано с помощью Datawrapper

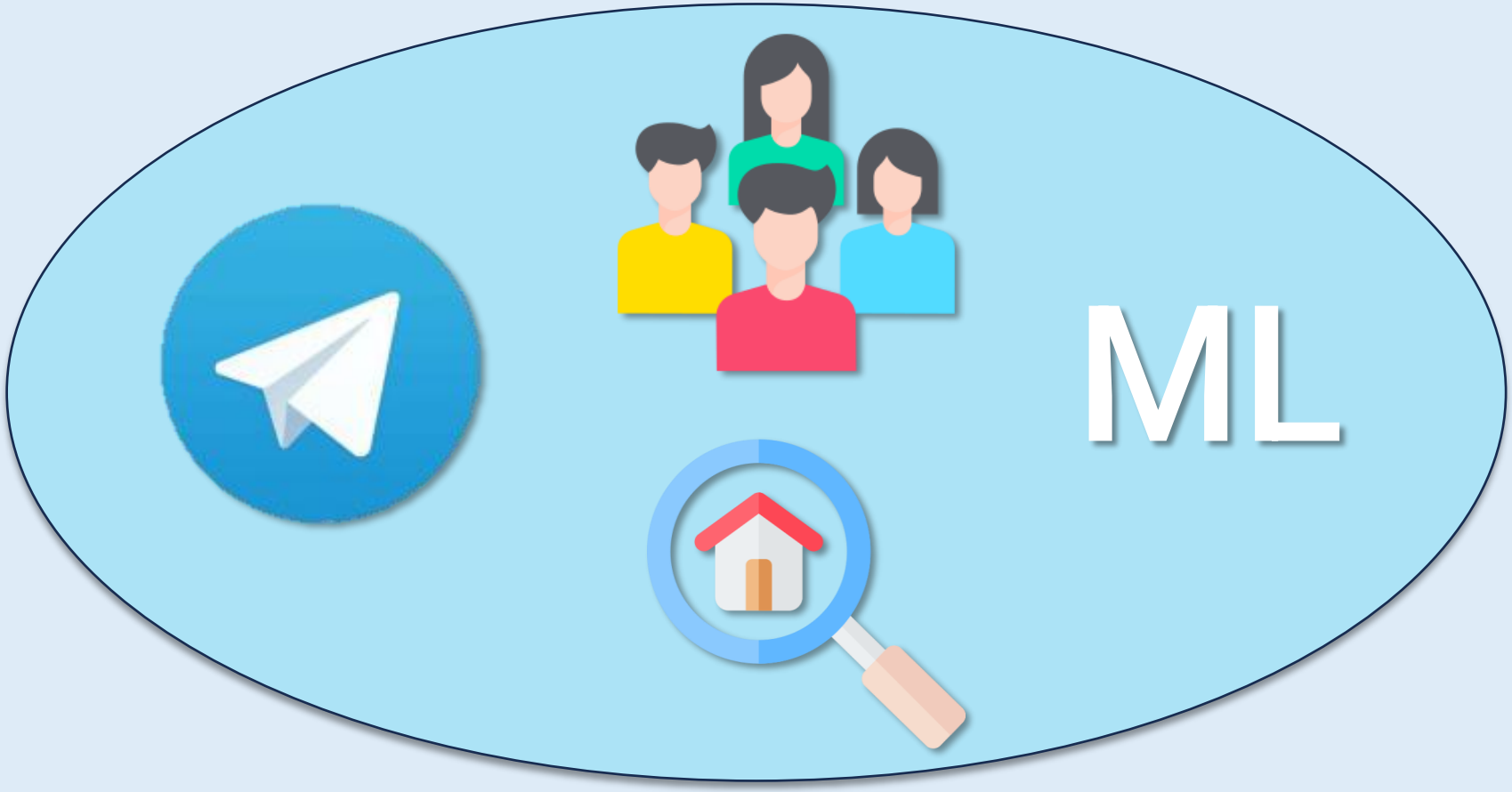
Machine learning

Machine Learning is a method of data processing and analysis that enables computers to use existing data to predict future behavior, outcomes, and trends. With machine learning, computers learn without being explicitly programmed. ML tools leverage artificial intelligence systems capable of identifying patterns and establishing connections through data experience.

Automated machine learning predictions can make applications and devices smarter.

Solution

«Apartment_rent_price_Bot»



Business goal

Business Goal:

To deliver a user-friendly and precise rental price estimation tool, driving user engagement and increasing the competitiveness of our real estate platform/service.

ML Project Goal

ML Project Goal:

To build a predictive model that estimates median rental prices using location (district), room count, and area features. The focus is on developing a robust algorithm with strong predictive accuracy to enhance our platform's competitiveness in the real estate sector.

Key elements preserved:

- Input parameters (district, rooms, area → price)
- Accuracy/reliability as core metrics
- Market value proposition

Data set preparation

1. Data collection from cian.ru using the cianparser library.
2. Data preprocessing for ML model training:
   - Removing redundant fields (ad poster, price per sq. m., contact phone, street, building number, district, etc.).
   - Merging city-specific datasets followed by city-based sorting.
   - Selecting one field as the target variable for the model to predict (rental price).

```python
import cianparser

data = cianparser.parse(
    deal_type="rent_long",
    accommodation_type="flat",
    location="Москва",
    rooms=(1, 2, 3, 4),
    start_page=55,
    end_page=100,
    is_saving_csv=True,
)
```

```python
additional_settings = {
    "is_by_homeowner": True,
    "min_price": 1000000,
    "max_price": 10000000,
    "have_balconies": True,
    "have_loggia": True,
    "min_house_year": 1990,
    "max_house_year": 2023,
    "min_floor": 3,
    "max_floor": 4,
    "min_total_floor": 5,
    "max_total_floor": 10,
    "house_material_type": 1,
    "metro": "Московский",
    "metro_station": "ВДНХ",
    "metro_foot_minute": 45,
    "sort_by": "price_from_min_to_max",
}
```

Data set

Unnamed: 0	author	author_type	link	city	deal_type	accommodation_type	floor	floors_count	rooms_count	total_meters	price_per_m2	price_per_m
0	KOVCHEG	real_estate_agent	https://www.cian.ru/rent/flat/294715265/	Москва	rent	flat	9	17	1	42.0	1309	55000
1	ID 438029	realtor	https://www.cian.ru/rent/flat/294909620/	Москва	rent	flat	18	27	1	42.0	1785	75000
2	ПИК-Аренда	real_estate_agent	https://www.cian.ru/rent/flat/294738678/	Москва	rent	flat	5	10	1	22.6	1716	38800
3	ID 18835544	official_representative	https://www.cian.ru/rent/flat/278216280/	Москва	rent	flat	7	12	2	90.0	6666	600000
4	Apple Real Estate	real_estate_agent	https://www.cian.ru/rent/flat/293989512/	Москва	rent	flat	4	8	4	223.0	3363	750000
...	...	...	...	...	...	...	...	...	...	...	...	...
386	ID 86319814	homeowner	https://ekb.cian.ru/rent/flat/295069300/	Екатеринбург	rent	flat	10	12	3	65.0	615	40000
387	Высота	real_estate_agent	https://ekb.cian.ru/rent/flat/294512032/	Екатеринбург	rent	flat	3	24	3	90.0	444	40000
388	ID 109885690	homeowner	https://ekb.cian.ru/rent/flat/294996012/	Екатеринбург	rent	flat	13	32	2	63.0	714	45000
389	ID 51315956	homeowner	https://ekb.cian.ru/rent/flat/295018030/	Екатеринбург	rent	flat	9	9	3	79.0	506	40000
390	ID 109679561	homeowner	https://ekb.cian.ru/rent/flat/294778273/	Екатеринбург	rent	flat	14	25	1	41.0	731	30000

	Unnamed: 0	city	floor	floors_count	rooms_count	total_meters	underground
0	0	Москва	9	17	1	42.0	Новые Черёмушки
1	1	Москва	18	27	1	42.0	Университет
2	2	Москва	5	10	1	22.6	Бунинская аллея
3	3	Москва	7	12	2	90.0	Охотный ряд
4	4	Москва	4	8	4	223.0	Парк Культуры
...	...	...	...	...	...	...	...
386	386	Екатеринбург	10	12	3	65.0	NaN
387	387	Екатеринбург	3	24	3	90.0	Площадь 1905 года
388	388	Екатеринбург	13	32	2	63.0	Ботаническая
389	389	Екатеринбург	9	9	3	79.0	Ботаническая
390	390	Екатеринбург	14	25	1	41.0	Ботаническая

Regression task

Rental price prediction is a regression task that can be effectively addressed using regression ML models. One suitable library for this task is CatBoost - an open-source gradient-boosted decision trees library developed by Yandex.

Technical Implementation:

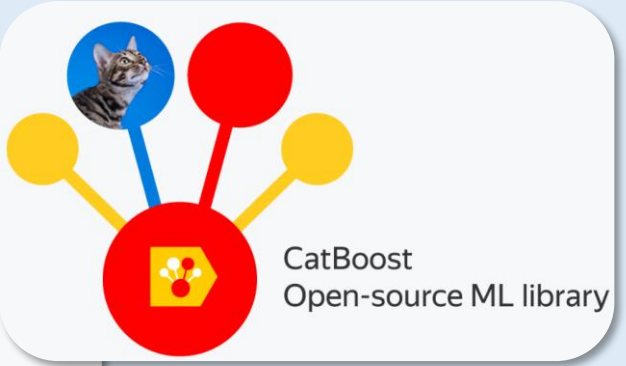
The method using decision trees as base algorithms is called Gradient Boosting over Decision Trees (GBDT)

GBDT demonstrates strong performance with:

- Tabular data structures
- Heterogeneous data types

ML-model

- **Model Architecture:** Gradient-boosted decision trees (*CatBoostRegressor* implementation).
- **Library:** CatBoost (Yandex-developed gradient boosting framework).
- **Features:**
  - Overfitting resistance via ordered boosting
  - Native support for missing data
  - Versatility (regression & classification)



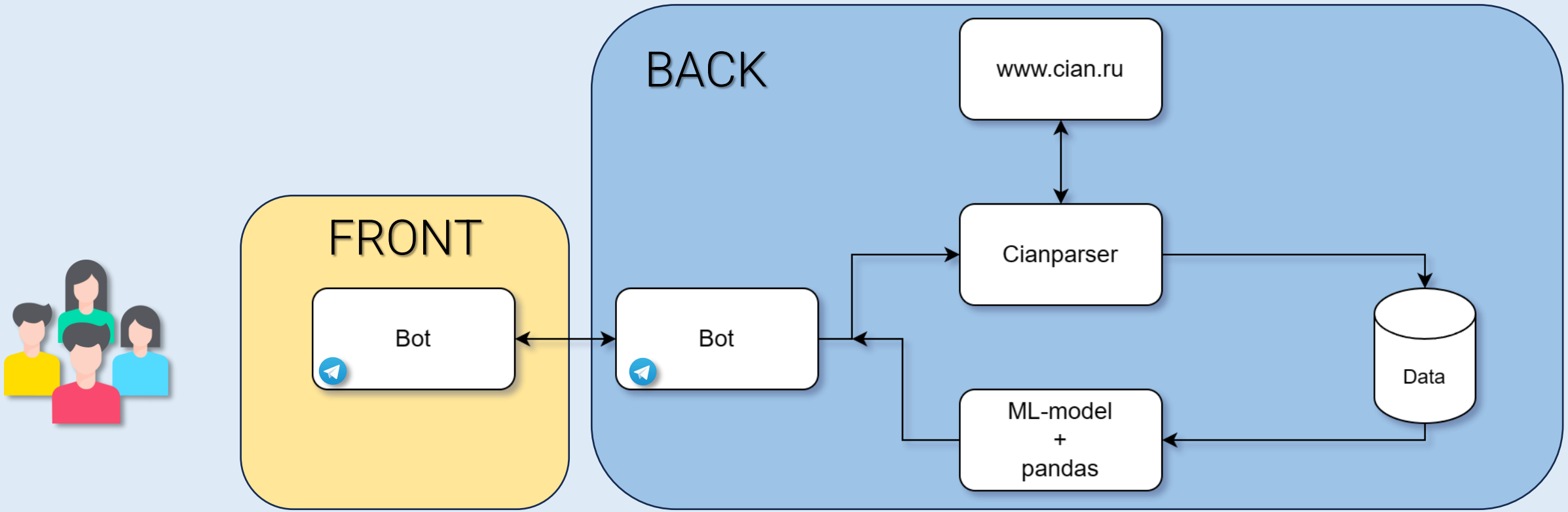

```
from catboost import CatBoostRegressor

model = CatBoostRegressor(iterations=100000,
                          learning_rate=0.7,
                          depth=3)
```

```
MAE = np.sum(np.abs(y_test - y_predict)) / len(y_test)
```

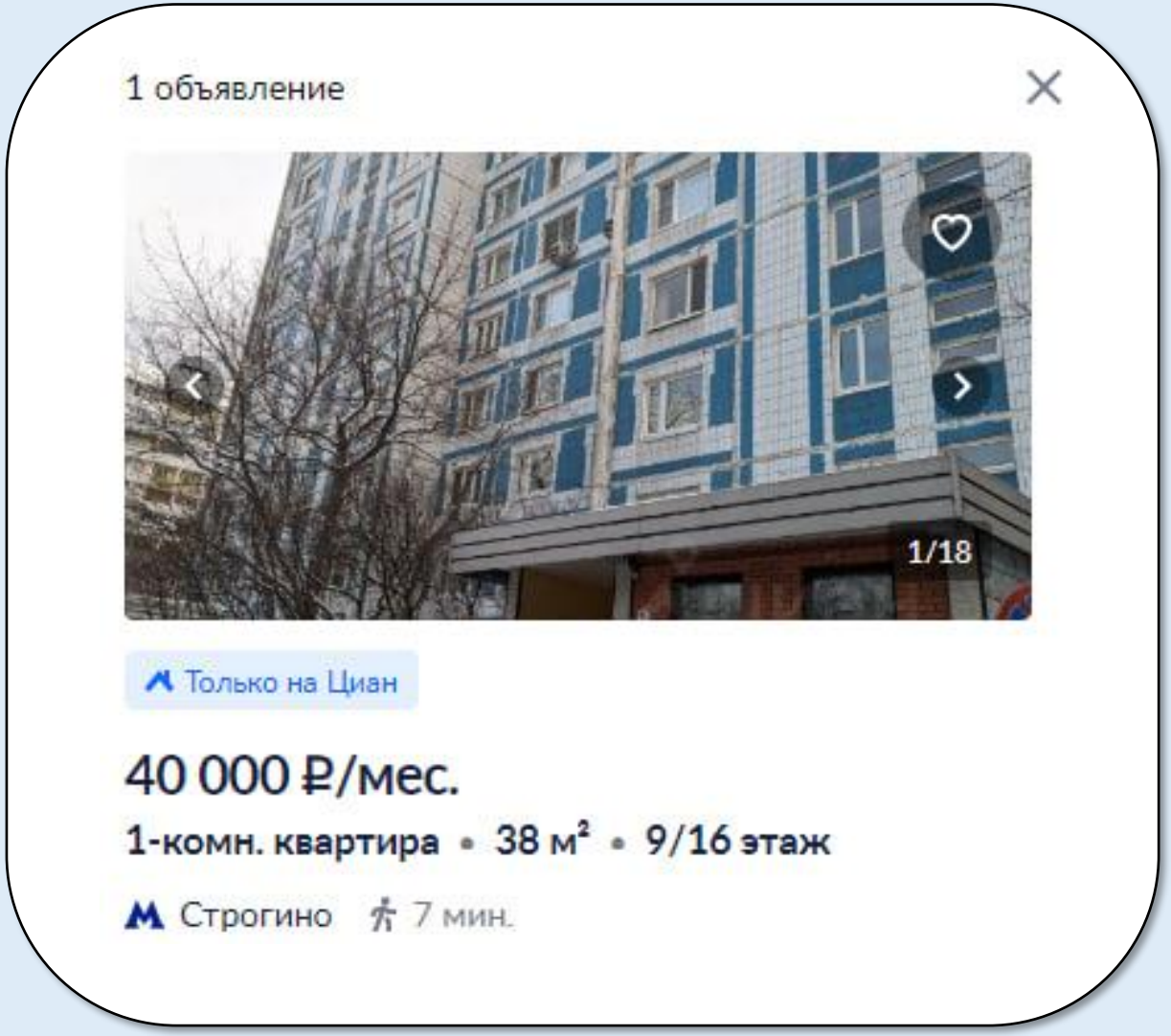
```
MAE
```

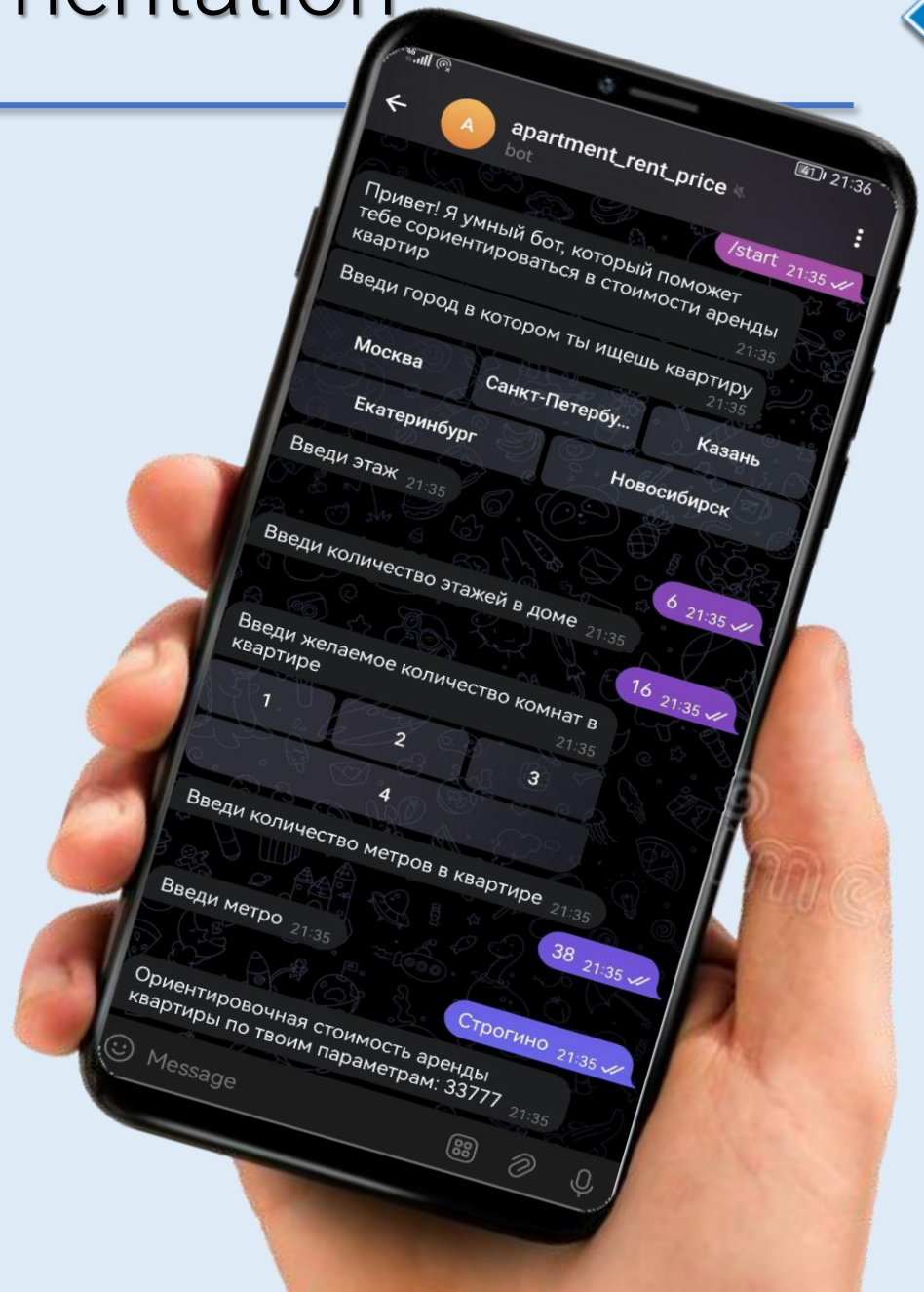
```
32374.373478242684
```

Assistant Architecture

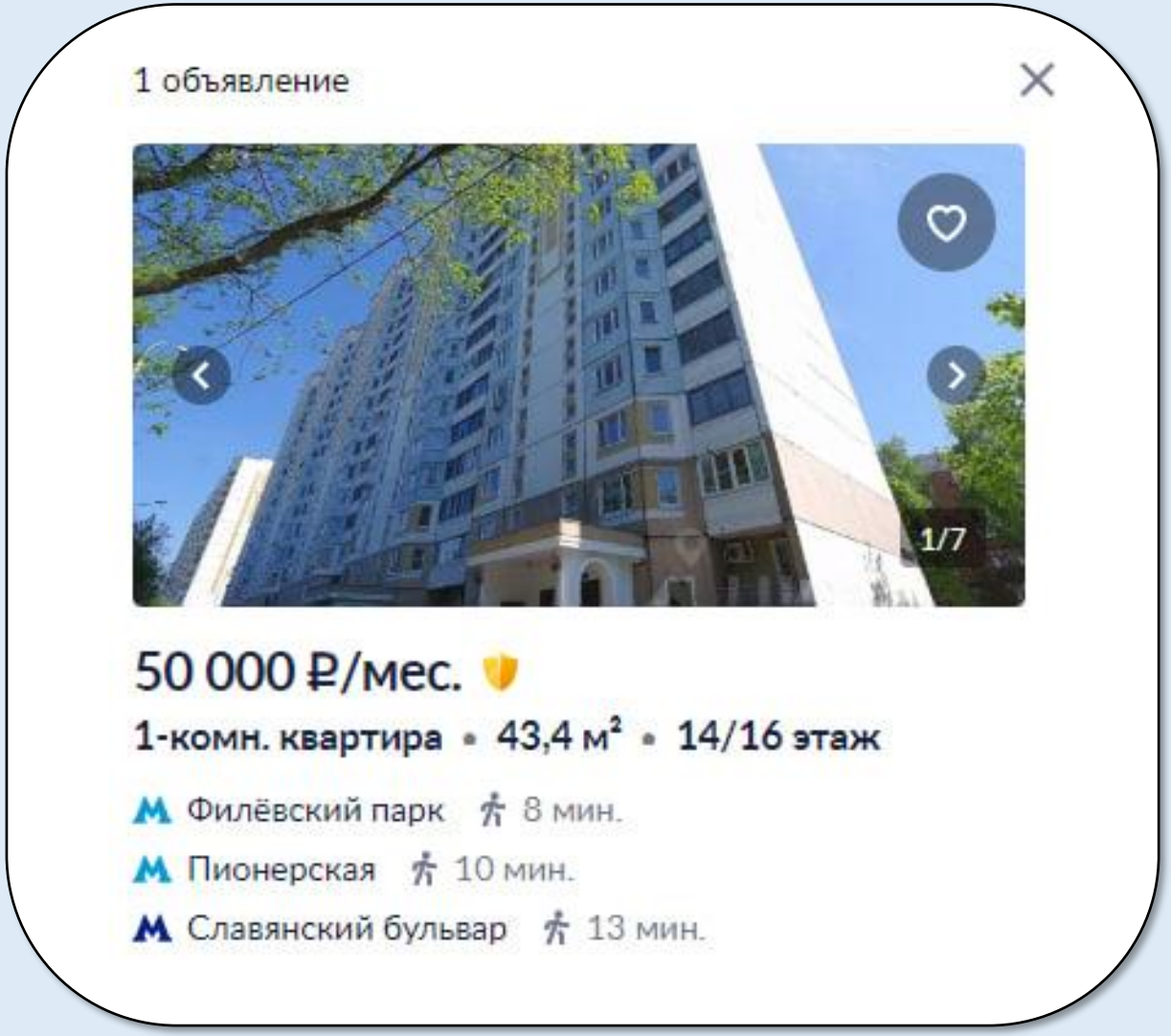


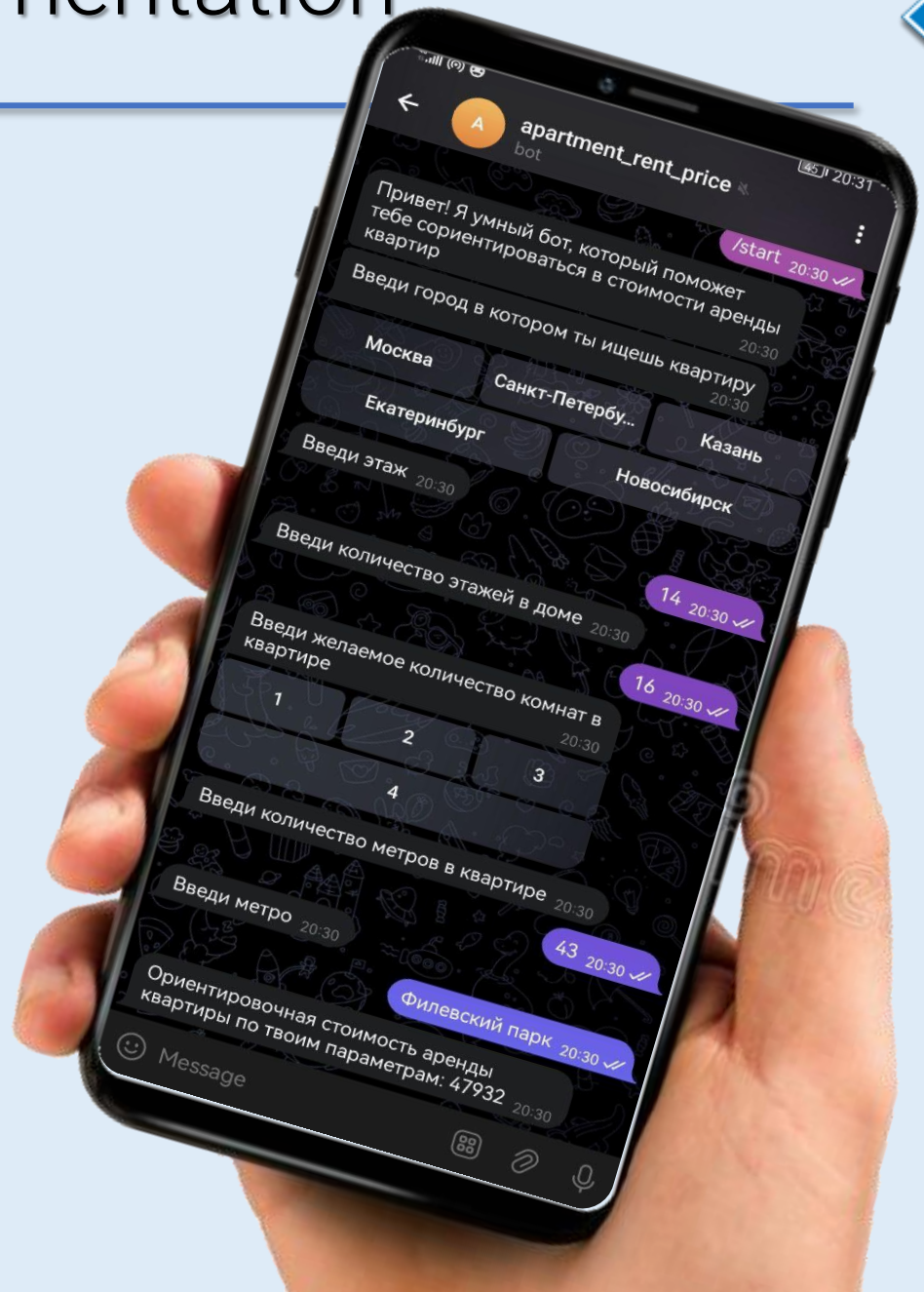
Results & Implementation



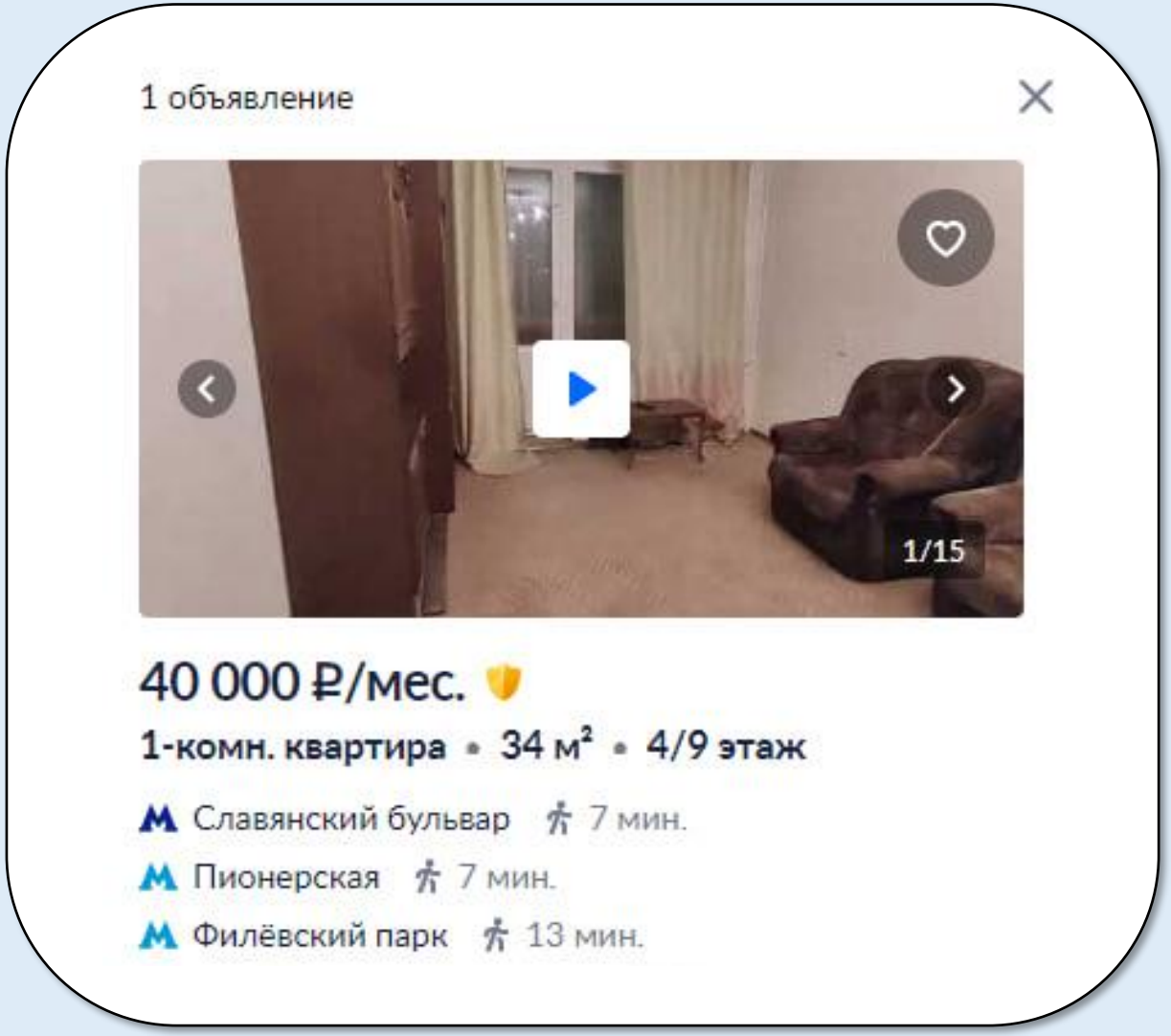


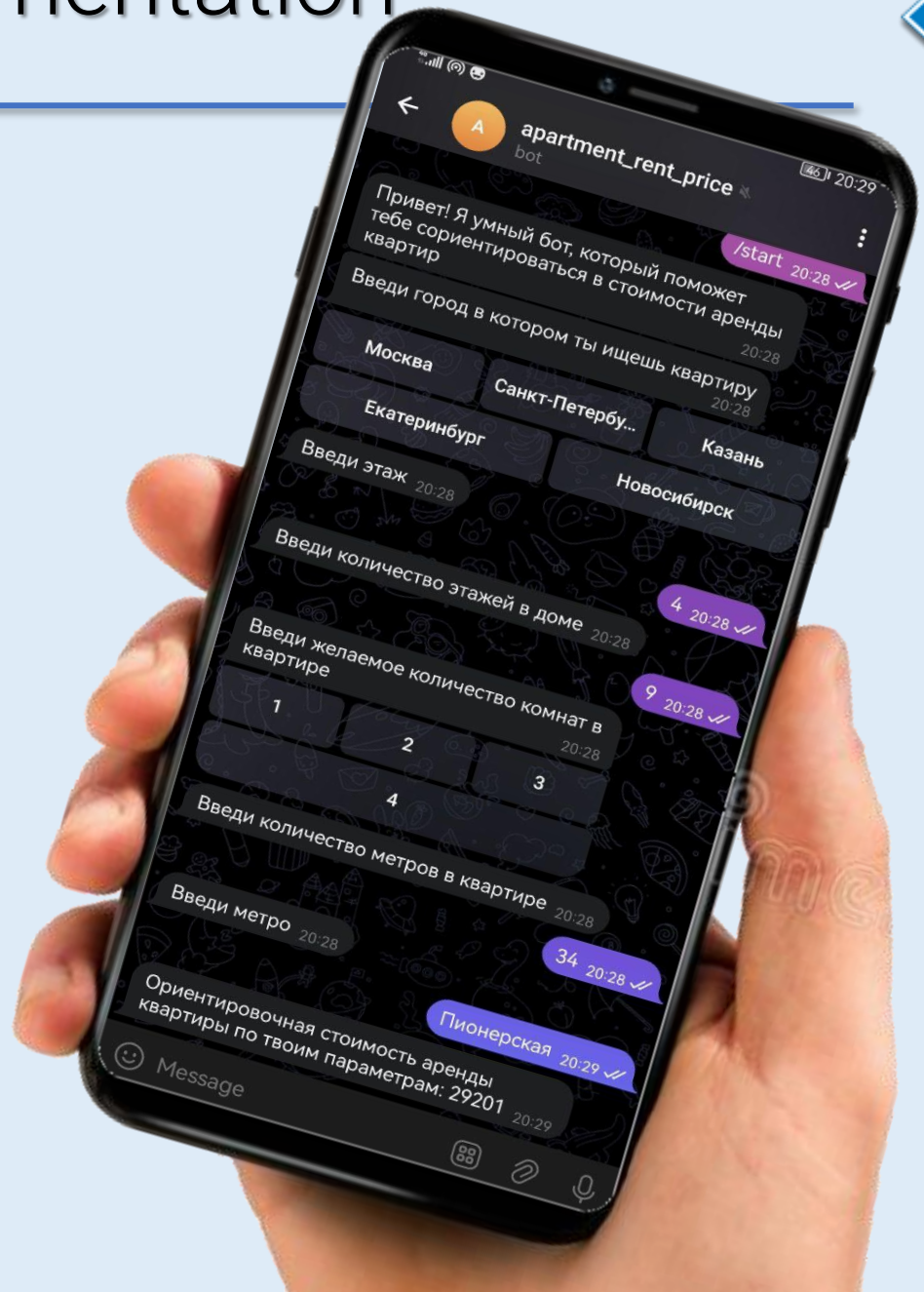
Results & Implementation



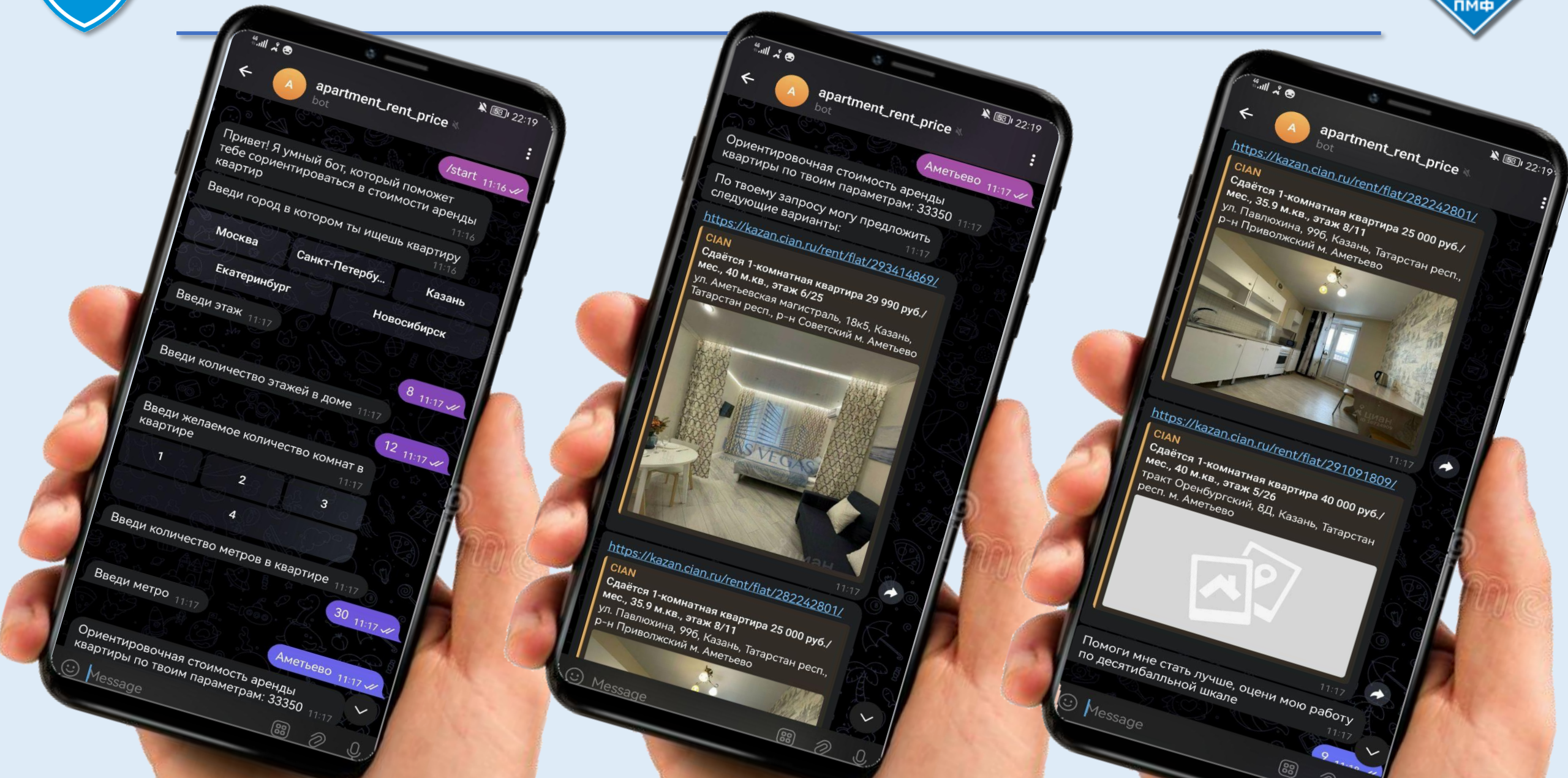


Results & Implementation

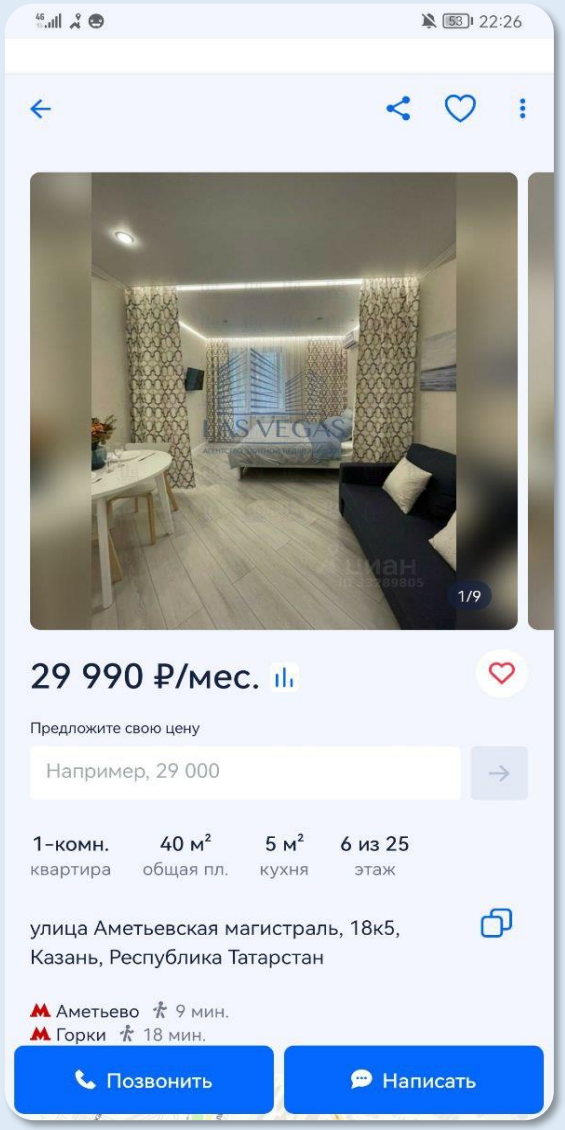


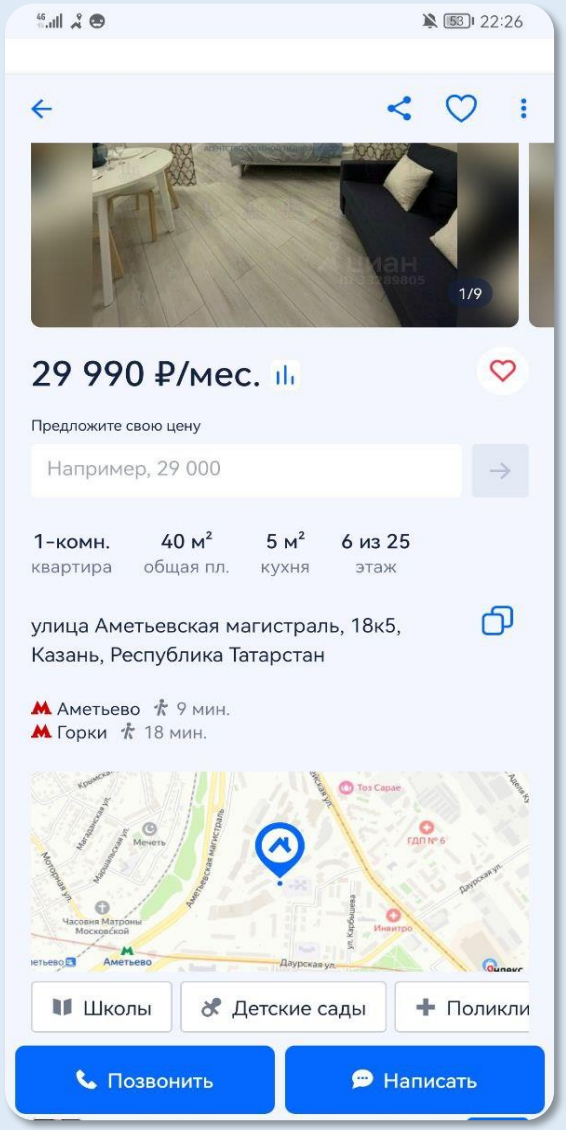


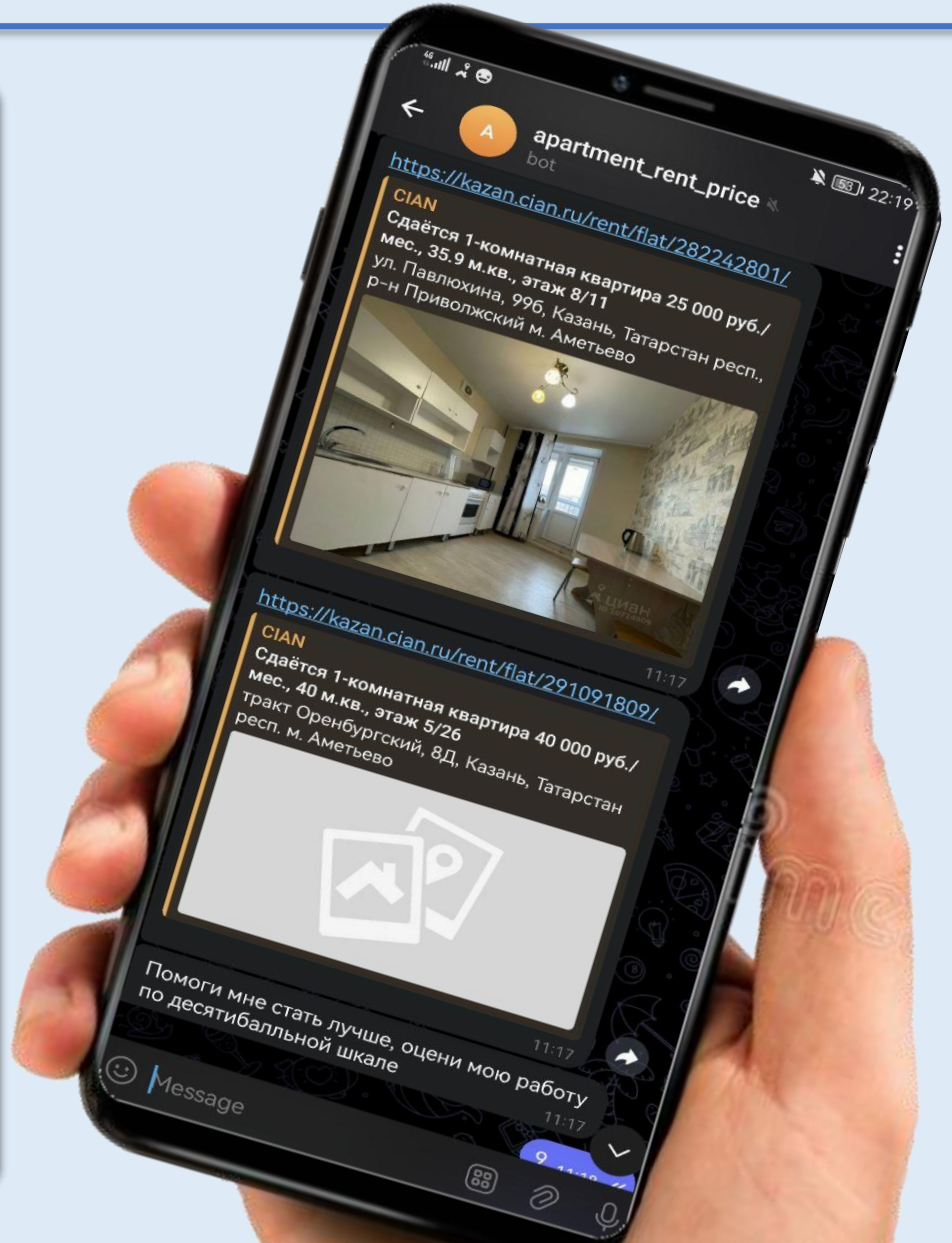
Results & Implementation



Results & Implementation







Conclusion

1. An analysis of potential ML application areas was conducted.
2. An analysis was performed to identify key challenges in the field, followed by the selection of one of them for an ML-based solution.
3. The necessary dataset was collected and processed.
4. A Telegram bot was implemented using appropriate libraries, serving as an intermediary between the client and the ML model.
5. An ML model was selected, trained, and deployed to predict the parameter of interest to the client.
6. The final output was achieved:
   - A parser and ML model gather the required dataset.
   - The most cost-effective rental price is calculated.
   - The most suitable option from the dataset is selected and recommended.

Thank you for attention!