

# Отчет по тестовому заданию

## Постановка задачи

Необходимо построить математическую модель предсказания стоимости вторичной жилой недвижимости в зависимости от параметров жилья, используя данные из объявлений по городу Магнитогорск.

Следует иметь в виду, что реальная стоимость жилья (при совершении сделки) как правило ниже стоимости, указанной в объявлении.

## Выбор и получение исходных данных

Исходные данные собираются с сайта [citystar.ru](https://citystar.ru) при помощи библиотеки BeautifulSoup. Для предсказания цены жилья используются следующие параметры:

- район города;
- улица города;
- этаж;
- количество этажей в доме;
- количество комнат;
- планировка;
- общая площадь;
- жилая площадь;
- площадь кухни.

## Выбор метода решения

Для решения задачи используются композиционные алгоритмы машинного обучения, основанные на деревьях решений. Данные алгоритмы не чувствительны к выбросам (кроме целевого признака), масштабу и линейной зависимости признаков, а некоторые из них способны работать с пропусками в данных. Это сильно снижает трудозатраты на предобработку данных, при том, что данные алгоритмы почти не уступают в качестве решения задачи алгоритмам на основе нейронных сетей.

## Описание алгоритма решения

Решение задачи состоит из следующих этапов:

- сбор данных с веб-сайта;
- предобработка данных;
- исследование данных (корреляция, пропуски, выбросы);
- подготовка данных для обучения и валидации моделей;
- обучение моделей;
- валидация моделей.

## Описание модели

Обучены следующие модели:

- случайный лес;
- градиентный бустинг на решающих деревьях;
- градиентный бустинг CatBoost.

Необходимо отметить, что ни для одной из моделей не производился подбор гиперпараметров (за исключением шага обучения и количества базовых алгоритмов).

## Описание качества полученных результатов

Для получения результатов работы моделей использовалась валидационная выборка и следующие метрики качества:

- средняя абсолютная ошибка (MAE);
- корень из среднеквадратичной ошибки (RMSE);
- средняя абсолютная ошибка в процентах (MAPE).

В таблице 1 приведены сравнительные результаты моделей.

Таблица 1 – Сравнительные результаты моделей на валидационной выборке

| Модель              | MAE, т. р. | RMSE, т. р. | MAPE, % |
|---------------------|------------|-------------|---------|
| Случайный лес       | 522        | 746         | 17,51   |
| Градиентный бустинг | 542        | 796         | 19,14   |
| CatBoost            | 522        | 706         | 17,54   |

Модель градиентного бустинга показала худший результат по всем представленным метрикам. Наиболее предпочтительной моделью можно считать CatBoost. Однако данная работа представляет из себя минимально жизнеспособный продукт (MVP), а потому, приведенные результаты могут значительно измениться при более тщательной обработке данных и подборе гиперпараметров.

## Описание результатов тестирования модели

Тестирование моделей проводилось на веб-сервисе [ivankud.com/docs](http://ivankud.com/docs).

## Выводы

Композиционные алгоритмы машинного обучения на основе деревьев решений показали приемлемое качество на задаче предсказания цены вторичной жилой недвижимости.

Обученная модель CatBoost имеет значение метрики MAPE на валидационной выборке 17,54 %.

Представленные модели способны дать лучшее качество при тщательной обработке данных и подборе гиперпараметров.