ИНСТИТУТ
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

Разработка и тестирование приложения для прогнозирования будущих событий с использованием машинного обучения в промышленности.

Цель данной работы:

Разработка и тестирование приложения для промышленности, основанного на машинном обучении (сфера строительства)

Предмет исследования:

Рынок вторичной недвижимости

Объект изучения:

Данные о продажах недвижимости за 2022 и 2023 года

Актуальность

УНИВЕРСИТЕТ ИННОПОЛИС

Проблема оценки реальной стоимости недвижимости является актуальной и важной в сфере рынка жилья по нескольким причинам:

- Агенты недвижимости могут полагаться на субъективный опыт или ограниченные данные, что может привести к ошибочной оценке стоимости квартиры. Неправильная оценка стоимости может привести к затягиванию процесса продажи из-за завышенной или заниженной цены.
- Точная и обоснованная оценка цены может повысить доверие клиентов к агенту недвижимости, укрепляя его репутацию как надежного специалиста.
- Инструмент на основе машинного обучения может анализировать рыночные тенденции и включать эту информацию в оценку стоимости, что особенно полезно в быстро меняющихся рыночных условиях.

Инструмент на основе машинного обучения для предсказания цены квартиры не только улучшает качество и скорость процесса оценки недвижимости, но и добавляет ценность для агентов, клиентов и инвесторов за счет повышения точности, объективности и надежности оценок

Задачи:

УНИВЕРСИТЕТ ИННОПОЛИС

- **Анализ существующих решений**
- **Запуск базовых моделей**
- **Обучить Оценка качества результата по релевантным для задачи метрикам**
- **Получение отчетов по результатам**

Информация о данных:

УНИВЕРСИТЕТ ИННОПОЛИС

Загруженный набор данных представляет собой анонимизированный срез информации по сделкам вторичной недвижимости за последний год. Ввиду больших отличий в полях различного типа недвижимости для данной работы были выбраны только квартиры на вторичном рынке. Для новостроек, домов, участков и дач требуется свой набор данных и своя модель, но сам алгоритм и обучение остается аналогичным.

price	price_per_meter	floor	rooms	repair_type	total_area	distance_to_center	district	house_build_year	house_type	balcony_number	bathroom_many	separate_bathroom	first_floor	geo_lat	geo_lon
17700000	159459	10	3	8	111.0	2018	Центр: Дом печати	2009	7	1	0	1	0	57.159912	65.562347
17200000	165385	6	3	7	104.0	1417	Центр: Драмтеатр	2012	4	0	1	1	0	57.151413	65.564637
17000000	159624	14	3	8	106.5	3450	Студгородок	2019	5	2	1	1	0	57.158096	65.596219
16950000	169500	7	3	8	100.0	3444	Тюменский-3	2016	7	0	0	1	0	57.113912	65.554611
16500000	165000	2	3	7	100.0	1313	Центр: Исторический	2009	7	0	0	1	0	57.156182	65.545880

Набор данных содержит 4842 строк и имеет 16 столбцов:

price (int) - сумма сделки по квартире
price per_meter (int) - цена за квадратный метр
floor (int) - этаж
rooms - количество комнат
repair_type (int) - тип ремонта согласно мастер справочнику
total_area (float) - общая площадь
distance_to_center (int) - расстояние до центра Тюмени в метрах.
district (string) - Район города

house_build_year (int) - год постройки здания
house_type (int) - тип стен здания, согласно мастер справочнику
balcony_number (int) - количество балконов
bathroom_many (int) - больше, чем одна ванная комната
separate_bathroom (int) - раздельный санузел
first_floor (int) - находится ли на первом этаже здания
geo_lat (float) - широта
geo_lon (float) - долгота

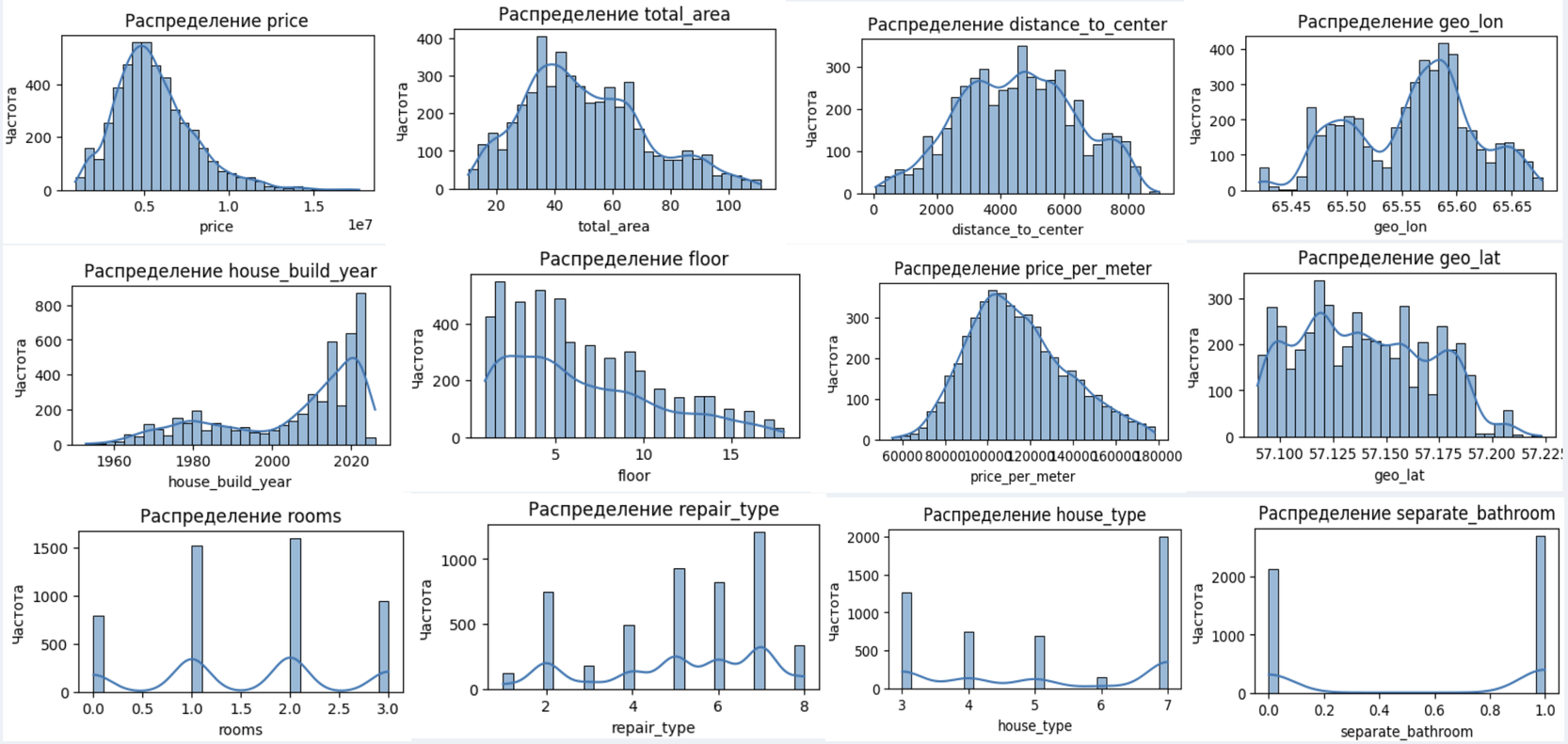
Распределение данных и типы столбцов

УНИВЕРСИТЕТ ИННОПОЛИС

	count	mean	std	min	10%	25%	50%	75%	90%	max
price	4842.0	5.606351e+06	2.320312e+06	945000.000000	3.050000e+06	4.000000e+06	5.299000e+06	6.800000e+06	8.600000e+06	1.770000e+07
price_per_meter	4842.0	1.142783e+05	2.321875e+04	55263.000000	8.649540e+04	9.774925e+04	1.113815e+05	1.290320e+05	1.470806e+05	1.779660e+05
floor	4842.0	6.464271e+00	4.282918e+00	1.000000	2.000000e+00	3.000000e+00	5.000000e+00	9.000000e+00	1.300000e+01	1.800000e+01
rooms	4842.0	1.553284e+00	9.803642e-01	0.000000	0.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00	3.000000e+00	3.000000e+00
repair_type	4842.0	5.137133e+00	1.956717e+00	1.000000	2.000000e+00	4.000000e+00	5.000000e+00	7.000000e+00	7.000000e+00	8.000000e+00
total_area	4842.0	5.043707e+01	2.111918e+01	10.400000	2.500000e+01	3.500000e+01	4.730000e+01	6.400000e+01	8.200000e+01	1.110000e+02
distance_to_center	4842.0	4.544238e+03	1.817122e+03	65.000000	2.238000e+03	3.181000e+03	4.594000e+03	5.828000e+03	7.126000e+03	8.974000e+03
house_build_year	4842.0	2.005427e+03	1.787991e+01	1953.000000	1.977000e+03	1.991000e+03	2.013000e+03	2.020000e+03	2.022000e+03	2.026000e+03
house_type	4842.0	5.177200e+00	1.688996e+00	3.000000	3.000000e+00	3.000000e+00	5.000000e+00	7.000000e+00	7.000000e+00	7.000000e+00
balcony_number	4842.0	7.368856e-01	5.163696e-01	0.000000	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00
bathroom_many	4842.0	8.240397e-02	2.750076e-01	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
separate_bathroom	4842.0	5.590665e-01	4.965502e-01	0.000000	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
first_floor	4842.0	8.839323e-02	2.838952e-01	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
geo_lat	4842.0	5.713924e+01	3.010816e-02	57.089433	5.709901e+01	5.711541e+01	5.713642e+01	5.716208e+01	5.718205e+01	5.722301e+01
geo_lon	4842.0	6.556028e+01	5.611105e-02	65.420808	6.547922e+01	6.551036e+01	6.556861e+01	6.559612e+01	6.563643e+01	6.567748e+01

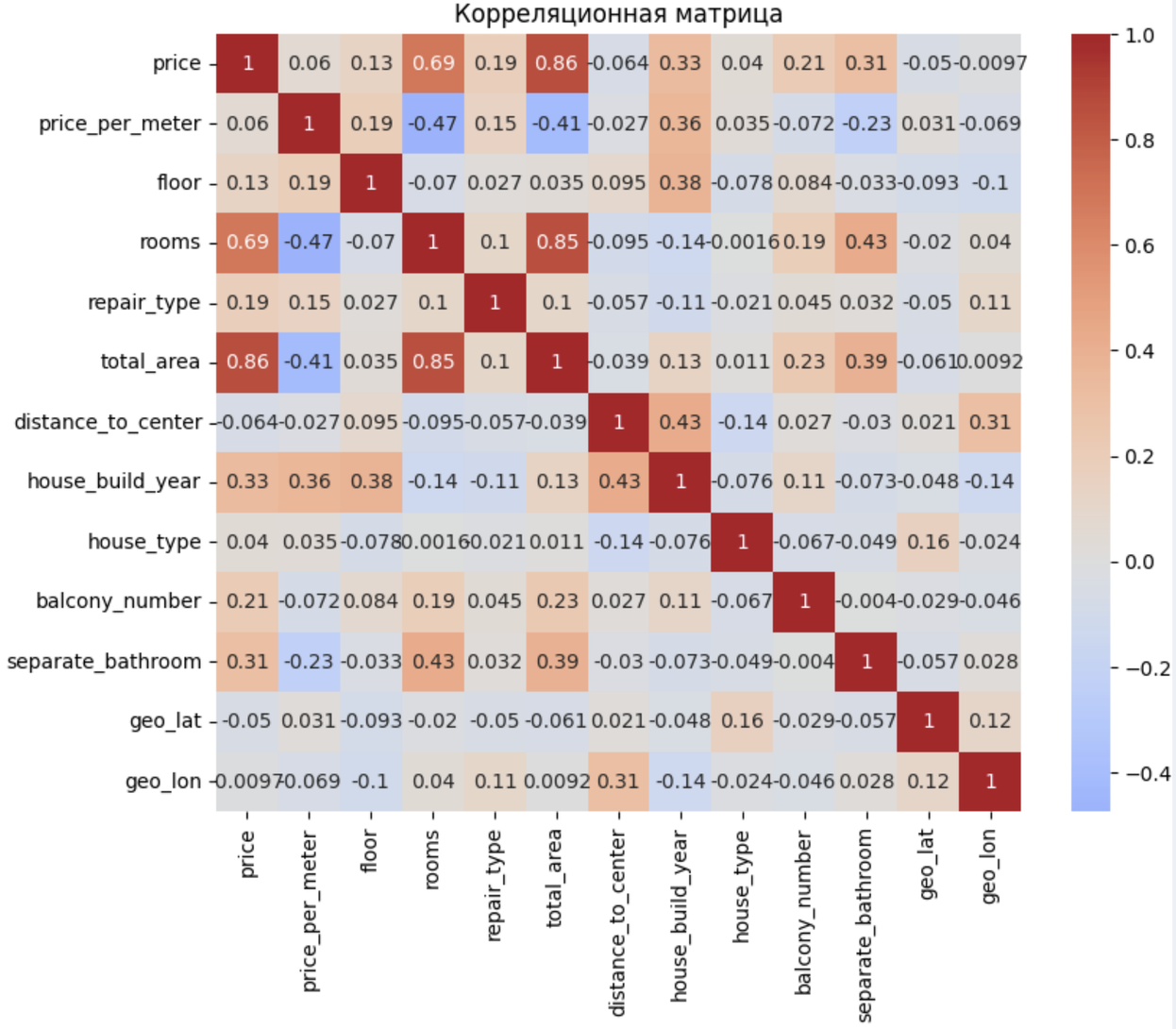
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4842 entries, 0 to 4841
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   price               4842 non-null   int64  
 1   price_per_meter     4842 non-null   int64  
 2   floor               4842 non-null   int64  
 3   rooms               4842 non-null   int64  
 4   repair_type         4842 non-null   int64  
 5   total_area          4842 non-null   float64
 6   distance_to_center  4842 non-null   int64  
 7   district            4842 non-null   object 
 8   house_build_year    4842 non-null   int64  
 9   house_type          4842 non-null   int64  
 10  balcony_number      4842 non-null   int64  
 11  bathroom_many       4842 non-null   int64  
 12  separate_bathroom   4842 non-null   int64  
 13  first_floor         4842 non-null   int64  
 14  geo_lat             4842 non-null   float64
 15  geo_lon             4842 non-null   float64
dtypes: float64(3), int64(12), object(1)
memory usage: 605.4+ KB
```

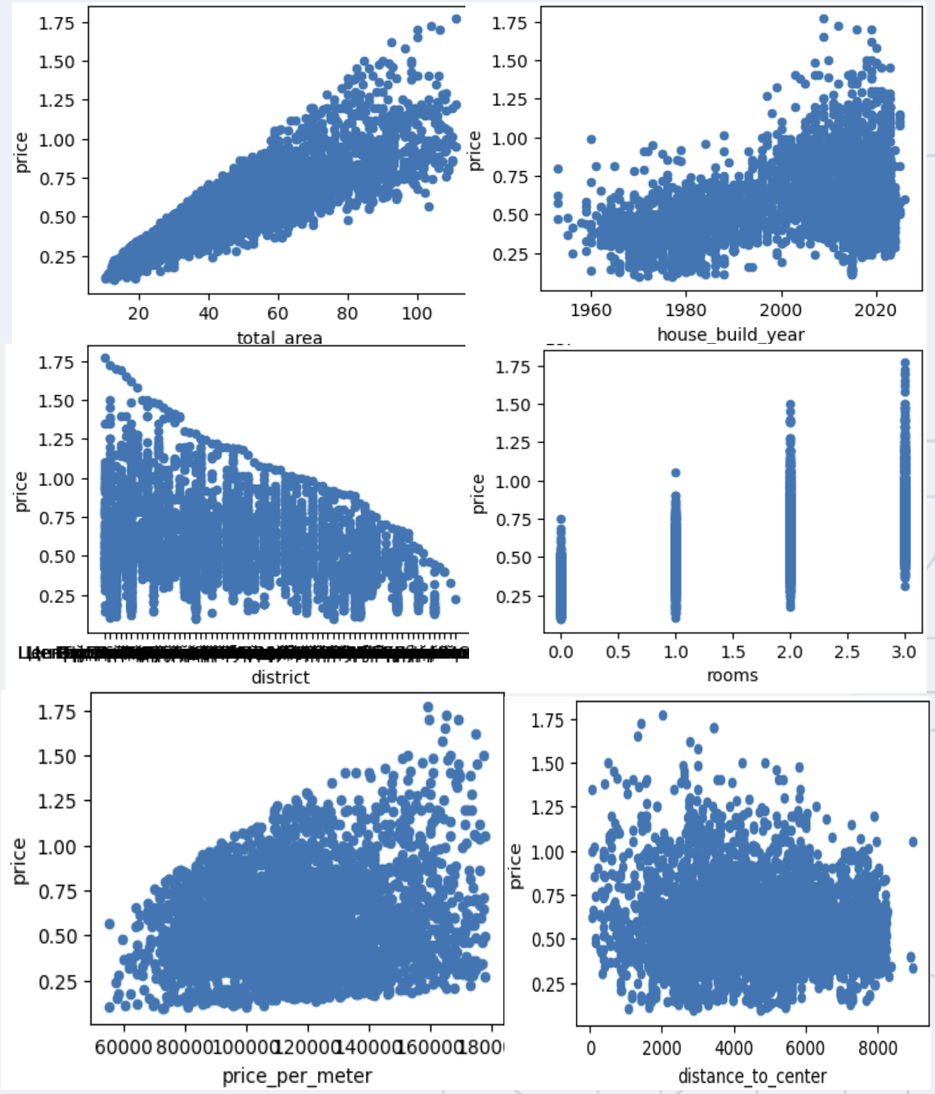
Визуализация данных



Матрица корреляции

УНИВЕРСИТЕТ ИННОПОЛИС





Подготовка данных

УНИВЕРСИТЕТ ИННОПОЛИС

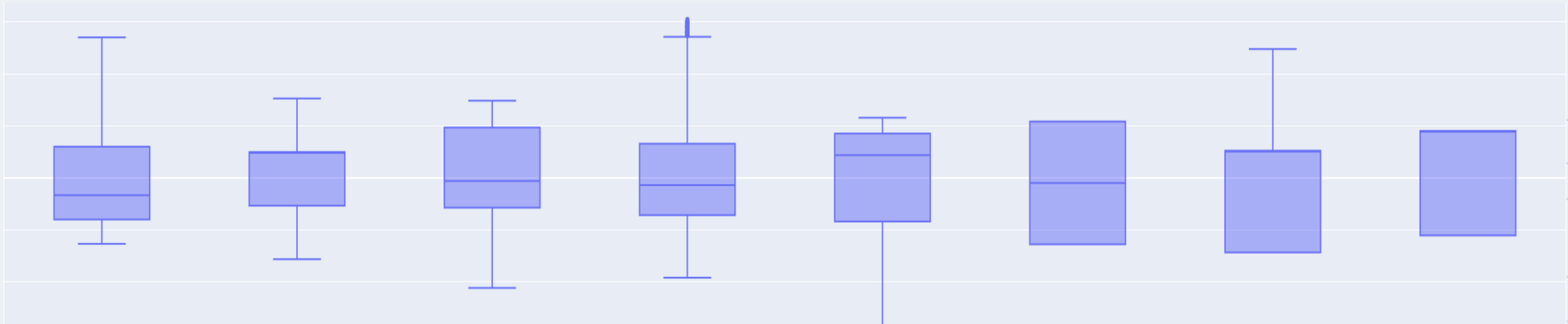
Удалены лишние столбцы, которые не будут принимать участия в обучении модели ввиду слабой корреляции или взаимной корреляции:
bathroom_many, first_floor, geo_lat,geo_lon, distance_to_center, price_per_meter
Количество строк после удаления выбросов: 4842

Для подготовки данных для обучения использован ColumnTransformer и Pipeline для удобной работы с данными. Итоговый препроцессор сохраняется также как и обученная модель.

К категориальным данным районов применен One-Hot Encoder, т.к. каждый из районов имеет свой вес и нам важно знать в каком именно районе находится квартира (BinaryEncoder не подходит), а также районы не имеют упорядоченности (LabelEncoder не подходит)

Т.к. после применения OneHotEncoding значения районов будут в диапазоне 0 -1, будет логично и остальные данные привести к этому диапазону, а значит MinMaxScaler предпочтительнее

Масштаб данных после удаления выбросов и нормализации с помощью MinMaxScaler:



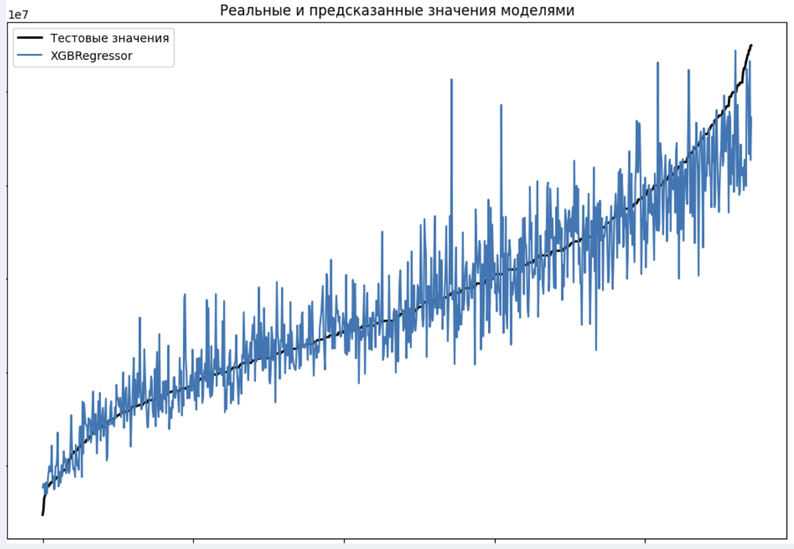
Машинное обучение:

УНИВЕРСИТЕТ ИННОПОЛИС

Для решения задачи произведено обучение пяти различных моделей, подходящих для решения задачи регрессии: Случайны лес, KNN, решающие деревья, Линейная регрессия, Градиентный бустинг.

По итогу модель градиентного бустинга показала наилучший результат и была выбрана в качестве основной

	model	r2	mse	rmse	mae
0	XGBRegressor	0.87	5.759906e+11	758940.0	542863.0
1	RandomForestRegressor	0.85	6.312866e+11	794535.0	581187.0
2	LinearRegressor	0.84	6.754141e+11	821836.0	605300.0
3	DecisionTreeRegressor	0.72	1.198181e+12	1094615.0	789170.0
4	KNeighborsRegressor	0.68	1.363257e+12	1167586.0	884113.0




```
# Подготавливаем данные для обучения модели
def makePreprocessor(data_frame):

  cat_features = ['district']
  num_features = list(data_frame.drop(cat_features, axis=1).columns)
  categorical_transformer = Pipeline(steps=[
      ('imputer', SimpleImputer(strategy='constant', fill_value='missing')),
      ('onehot', OneHotEncoder(handle_unknown='ignore'))])
  numeric_transformer = Pipeline(steps=[
      ('imputer', SimpleImputer(strategy='mean')),
      ('scaler', MinMaxScaler())])
  prepr = ColumnTransformer(
      transformers=[
          ('num', numeric_transformer, num_features),
          ('cat', categorical_transformer, cat_features)])
  return prepr
```

```
preprocessor = makePreprocessor(features)
features_preprocessed = preprocessor.fit_transform(features)
```

```
feature_names = preprocessor.get_feature_names_out()
```

```
# Разделение данных на обучающую и тестовую выборки
X = features_preprocessed
y = df[target_col]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, shuffle=True)
```

```
!pip install xgboost
import xgboost as xgb
xgb_model = xgb.XGBRegressor(objective='reg:squarederror')
xgb_model.fit(X_train, y_train)
```

Заключение:

УНИВЕРСИТЕТ ИННОПОЛИС

В результате анализа и обучения удалось получить модель, которая показывает хорошие метрики качества (R2 = 0.87) и вполне может быть ещё улучшена в будущем. Будут созданы отдельные модели для разных типов недвижимости, а также использованы макроэкономические параметры, чтобы модель лучше адаптировалась под внешнее влияние (цена марки brent, индекс CPI, ключевая ставка). Проект будет представлять собой RestApi сервис прогнозирования, а также сервис обновления и версионирования моделей.

Список литературы:

1. И.А.Дробинина «Изучение тенденции изменения стоимости одного квадратного метра квартиры на первичном рынке жилья» // Международный научно-технический журнал «Теория. Практика. Инновации». 2021. №1. С.26–30.
2. Н.В.Концевая «О моделировании рынка недвижимости и возможности прогнозирования цены квадрата» // Экономика, Статистика и Информатика. 2022. №4. С.31–34
3. Машинное обучение и анализ рынка недвижимости. «https://dzen.ru/a/XsX3Nk-RH3zNk2NU.
4. Строим свое будущее: как выбрать квартиру, опираясь на методы регрессионного анализа? https://habr.com/ru/articles/710000/

ИНСТИТУТ
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

Спасибо
за внимание