Команда 18. Сервис для предсказания стоимости недвижимости

Состав команды:

- Куликов Сергей
- Павлов Игорь
- Щербатюк Роман
- Тодоров Дмитрий

Куратор:

- Тимур Ермешев

Предобработка

- Очистка и фильтрация данных:
  - Удаление служебных столбцов (например, ссылка, Source_File, Sheet_Name)
  - Удаление дубликатов
  - Фильтрация аномальных значений (например, квартир с непропорционально большой площадью, этажей сверх разумного и т.п.)

- Обработка пропусков:
  - Заполнение пропущенных значений в числовых признаках с помощью KNNImputer

- Кодирование категориальных признаков:
  - Применение One-Hot Encoding для признаков с малым кол-вом уник значений
  - Объединение нескольких категориальных признаков в один (через функцию combine_factors) и последующее кодирование с помощью MultiLabelBinarizer
  - Остальные для CatBoost нет необходимости кодировать - подали на вход заполненные числовые признаки и категориальные с NaN значениями

Тестирование новых нелинейных моделей

- R2 - как основа для сравнения разных моделей
- RMSE - ошибка в рублях

Модель	Гиперпараметры	R^2	RMSE	MAPE %
Ridge base	alpha=10, max_iter=1000	0.7924	4786183	-
CatBoost	iterations=1000, learning_rate=0.1, depth=6	0.9372	3919376	9.03
CatBoost	подобранные Optuna: `'iterations': 1523, 'depth': 10, 'learning_rate': 0.09097382170808721, 'l2_leaf_reg': 0.025072850240151708, 'border_count': 135, 'random_strength': 0.004807066180128308, 'bagging_temperature': 0.009631257498284598`	0.9471	3598033	7.25

Модель	Гиперпараметры	R^2	RMSE	MAPE %
XGBoost	Default `'objective': 'reg:squarederror','max_depth': 10, 'eta': 0.1,'subsample': 0.8, 'colsample_bytree': 0.8,'eval_metric': 'rmse'`	0.9327	4 059 006	10.33
XGBoost	подобранные Optuna: `'max_depth': 12, 'eta': 0.052334147151371725, 'subsample': 0.947986259109503, 'colsample_bytree': 0.531758248415418, 'lambda': 0.4626453641183538, 'alpha': 0.01346128722153246`	0.9325	4 063 862	9.39
LightGBM	подобранные Optuna: `'max_depth': 13, 'learning_rate': 0.06917448221971932, 'num_leaves': 1789, 'feature_fraction': 0.7937032651952263, 'bagging_fraction': 0.9060713837551587, 'bagging_freq': 2, 'lambda_l1': 0.0009475875628289889, 'lambda_l2': 0.22937084043035164`	0.9470	3 602 581	7.30
Дерево решений	подобранные: `'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10`	0.9261	2 984 693	6.35
Случайный лес	Default	0.9603	2 188 182	4.76