

Задача Регрессии

RMSE: 0.1272

Место: 1532 из 5760

АЛЕКСЕЕВ АНДРЕЙ

SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmntSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00
	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmntSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt

HOUSE PRICES ADVANCED REGRESSION TECHNIQUES

DATASET



House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

 kaggle



ЦЕЛЬ: ПРЕДСКАЗАТЬ ЦЕНУ НА
НЕДВИЖИМОСТЬ И ПОЛУЧИТЬ
МИНИМАЛЬНЫЙ **RMSE**

ИТОГОВОЕ РЕШЕНИЕ

Обработка данных	Обработка таргета	Модель
Кодирование категориальных признаков, нормализация признаков и преобразование Бокса-Кокса	Логарифмирование предсказываемого значения	Бустинг с помощью CatBoostRegressor

Работа с признаками

ЗАПОЛНЕНИЕ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

Многие фичи в датасете были пропущены. Вместо удаления данных столбцов было предложено заполнить их None или 0

БЫЛ ИСПОЛЬЗОВАН LABELENCODER

Данный подход показал лучшее качество, чем LabelBinarizer при обучении моделей (0.12 против 0.14)

БЫЛ ИСПОЛЬЗОВАН ROBUSTSCALER

Такая нормализация фичей позволяет привести датасет к нулевому среднему и единичной дисперсии без учета выбросов, тк для расчета статистик берутся данные между 1 и 3 квантилями

ИСПОЛЬЗОВАЛОСЬ ПРЕОБРАЗОВАНИЕ БОКСА-КОКСА

Использование данного преобразования помогло также улучшить результат регрессии

ДОБАВЛЕНА ФИЧА, КОТОРАЯ ВЫРАЖАЕТ ОБЩУЮ ПЛОЩАДЬ

Фича выражает собой общую площадь здания

Опробованные методы

Метод	Lasso (линейный)	ElasticNet (линейный)	KNN (метрический)	CatBoost (бустинг)	RandomForest
Настраиваемые параметры	-//-	-//-	-//-	-//-	-//-
Область настройки	-//-	-//-	-//-	-//-	-//-
Результат на кросс- валидации (выбран лучший)	0.1350	0.1350	0.2063	<u>0.1207</u>	0.1456
Результат на лидерборде (был отправлен только лучший алгоритм)	-//-	-//-	-//-	0.1272 (место 1532 из 5760)	-//-

*Также был опробован подход с усреднением предсказаний моделей, но он не улучшил качество по сравнению с Catboost