USED CARS PRICE PRIDICATION

Tasneem Yousef Halawani
SDAIA Data Science Bootcamp

PROBLEM

Predicting used cars prices in Saudi Arabia

DATA

USED CARS PRICES AND SPEC SCRAPED FROM SYARAH WEBSITE

- Acquired from: https://www.kaggle.com/turkibintalib/saudi-arabia-used-cars-dataset
- Updated 5 months ago
- 8,248 records of used cars
- 15 features:

Link to the car page	the color of the car	its condition
brand name	options	the covered mileage
model	capacity of the engine	region
manufacturing year	type of fuel	price
origin	transmission type	negotiable

DATASET PREPROCESSING

* Dropped unneeded columns: **Link** and **Condition**

* Dropped rows with undefined (negotiable) price

* Converted **Price** column to float

* Dropped NAs

* Added **Province** column

* Checked the outliers in all numerical columns, but didn’t change them

* Visualized the correlation between all attributes

* Converted categorical features into numerical by getting dummies

MODELS

* Ended up with 4,404 rows and 476 columns

* Target: **Price , all** other attributes used as independent variables

* Standard Scale the independent attributes

* training set size = **80%**

MODELS

	LinearRegression	GradientBoosting Regressor	XGBRegressor	LGBMRegressor
Mean absolute error	6.014811170313652e+17	17667.08	16116.86	18814.41
Mean squared error	2.173336009163085 3e+37	1525865539.93	1167669525.89	1967576255.71
Root Mean squared error	4.661905199768744e+18	39062.33	34171.18	44357.37
Median absolute error	12520.0	9049.94	8644.36	8930.38
Explain variance score	-3.005927248569663e+27	0.79	0.84	0.73
R2 score	-3.019865590866434e+27	0.79	0.84	0.73

MODELS

DROPPED: 'TYPE','CITY','YEAR'

	LinearRegression	GradientBoosting Regressor	XGBRegressor	LGBMRegressor
Mean absolute error	2.344747303625326e+16	24453.49	24229.61	25195.39
Mean squared error	1.6227967135385517e+35	2304968941.63	2252585841.39	2324787831.52
Root Mean squared error	4.0283951066628896e+17	48010.09	47461.41	48216.05
Median absolute error	20490.0	13626.05	13072.29	13987.44
Explain variance score	-2.601537354555921e+25	0.63	0.64	0.63
R2 score	-2.610381011191642e+25	0.63	0.64	0.63

CONCLUSION

* including all attributes yeilded better results

* in both cases **XGBRegressor** resulted in better accuracy with all paramaters set to default except: (objective ='reg:linear', max_depth=5, n_estimators = 100)

* having larger dataset would help getting better results

THANK YOU

any questions?