HOUSE PRICE PREDICTION SYSTEM

Ozgur Lezgiyev

PROBLEM

Real estate is a very important pillar of the economy. However, despite a large amount of data available, we do not have precise measurements of house prices.

Therefore, this study aims to apply machine learning to predict home selling prices based on various economic indicators.

INTRODUCTION

The real estate industry is one of the most competitive in terms of pricing and is always changing. It is one of the key areas where machine learning concepts are applied to improve and accurately predict expenses. Predicting a real estate property's market worth is the paper's main goal. This approach aids in determining a property's beginning price depending on geographic factors. Future costs will be predicted by dissecting past market trends, price ranges, and upcoming technological improvements. This examination means to predict house prices in Texas with Random Forest regressor. It will help clients to put resources into a bequest without moving toward a broker. The result of this research proved that the Random Forest regressor gives an accuracy of 96.8%.

DATA

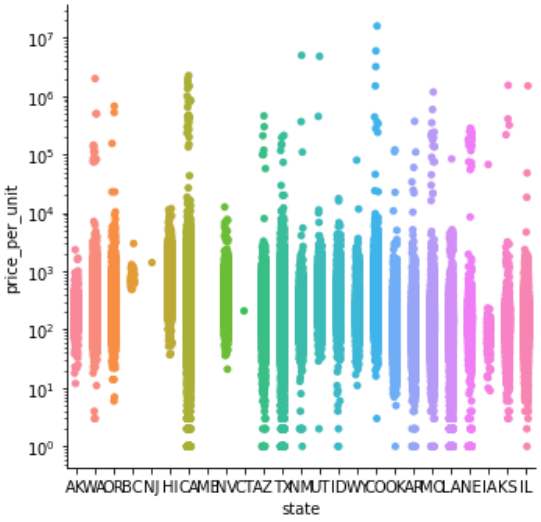
The Data used for this project consists of 600,000 rows by 1547 features, including house prices from all over the US, which are listed at Zillow.com. A large number of dataset samples makes it a "tall" dataset which will make it possible to achieve useful predictive accuracy for a wide range of locations and home types. The first step is to clean this data and ensure fits for purpose. Then look at which features would provide interesting insights and answer questions that would help to predict the value of a house. Finally, dive into building a model capable of predicting house prices.

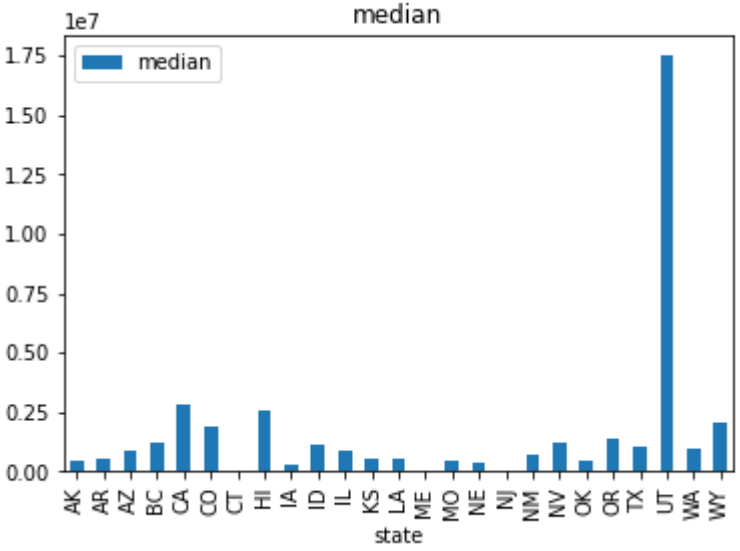
	property_	property_	address	street_na	apartmen	city	state	latitude	longitude	postcode	price	bedroom_	bathroom	price_per_
count	600000	600000	600000	599869	14815	599999	599999	529122	529122	599970	600000	443845	471733	435365
unique	600000		598588	339224	2664	7977	25			10820				
top	https://www.zillow.		(undisclos	(undisclos	1	Chicago	TX			84043				
freq	1		36	1713	312	14138	146636			1102				
mean		8.89E+08						36.28238	-105.814		605187	3.411964	2.471116	484.0448
std		9.72E+08						5.673355	13.46463		3175555	22.4916	5.684972	29855.27
min		27						18.98514	-165.408		0	0	0	0
0.25		54021427						32.61211	-117.346		170000	3	2	155
0.5		2.07E+08						35.40357	-101.897		369900	3	2	220
0.75		2.07E+09						39.66167	-95.3542		625000	4	3	331
max		2.15E+09						71.29917	-87.5256		2.15E+09	13210	1892	15900000

	living_spa	land_spac	land_spac	broker_id	property_	property_	year_build	total_num	listing_ag	RunDate	agency_na	agent_nar	agent_pho	is_owned_
count	447847	515119	515119	0	600000	600000	0	0	600000	600000	444524	0	0	600000
unique			2		7	2				1	34372			
top			acres		SINGLE_F/	FOR_SALE				44675.32	Coldwell Banker Realty			
freq			269149		354366	383365				600000	5936			
mean	4298.387	3099.427							-1					0.000498
std	178020.1	5350.082							0					0.022318
min	0	-13068							-1					0
0.25	1360	1.01							-1					0
0.5	1863	60							-1					0
0.75	2574	6534							-1					0
max	59913270	1746756							-1					1

The dataset has 1 table with a total of 16 categorical features 10 numerical features and 2 special features. This makes it a good dataset for learning how to extract features from location and DateTime features for spatiotemporal modeling.

Exploratory Data Analysis (EDA)





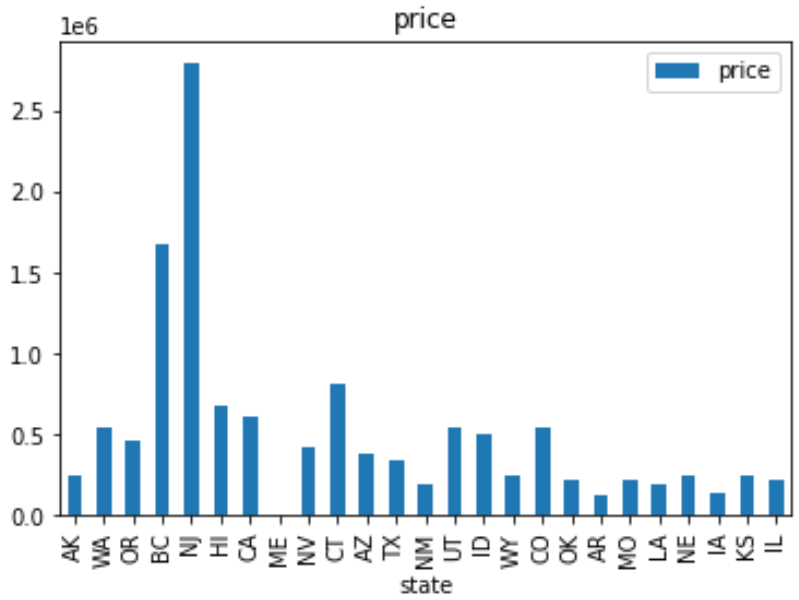
In the graphics on the upper left, you can see the distribution of prices and on the lower left standard deviation of prices by states

House prices in the state of California vary widely, but outliers draw attention, but in mid-western states such as Utah, Wyoming, and Colorado, base prices start higher

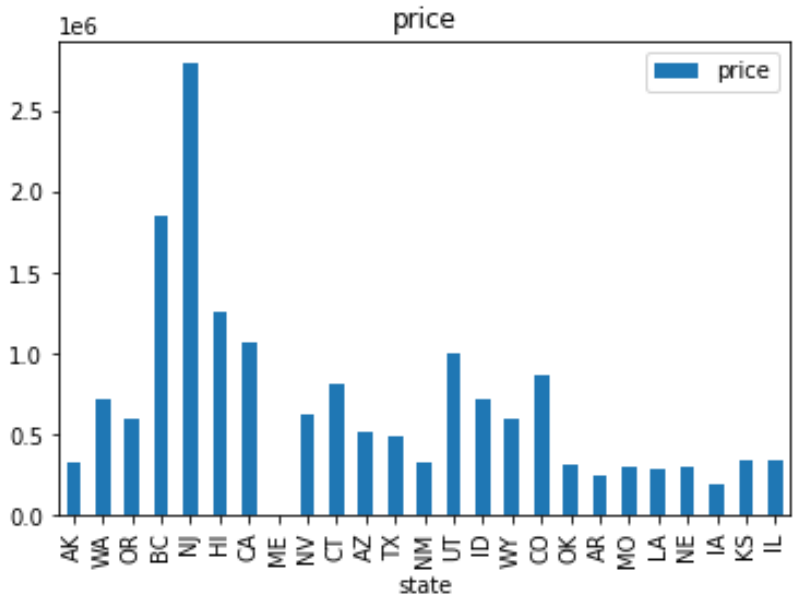
A more homogeneous distribution is seen in southern states such as Texas and Arizona

Moreover there is a clear sign of high standard deviation in Utah.

Median Prices



Mean Prices

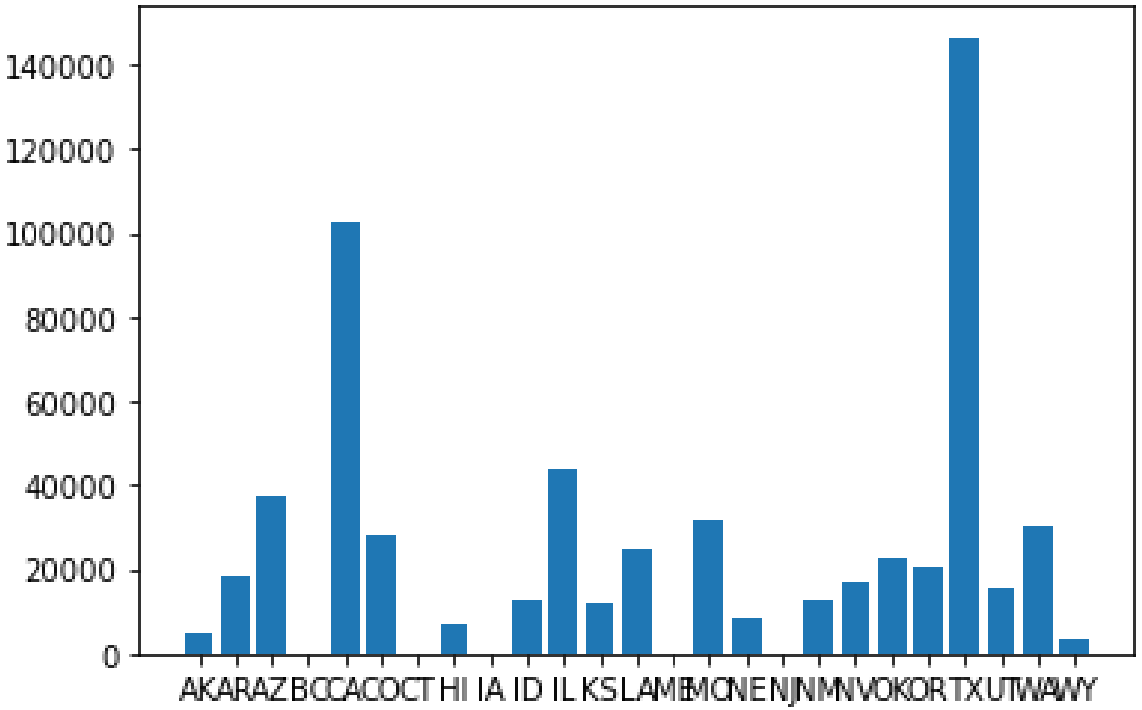


You can see the median price chart according to the states in the left chart, and the average price chart by the states in the right chart.

It is noticeable that there is no difference between the median value and the average value in states such as New Jersey and Connecticut, while in more western states such as Washington, Oregon, Utah, Idaho, Wyoming, and Colorado the average prices are significantly higher than the median prices, visible. From this point of view, we can deduce that there are more high-priced luxury houses in these states, and this drags the average values up.

Feature Selection

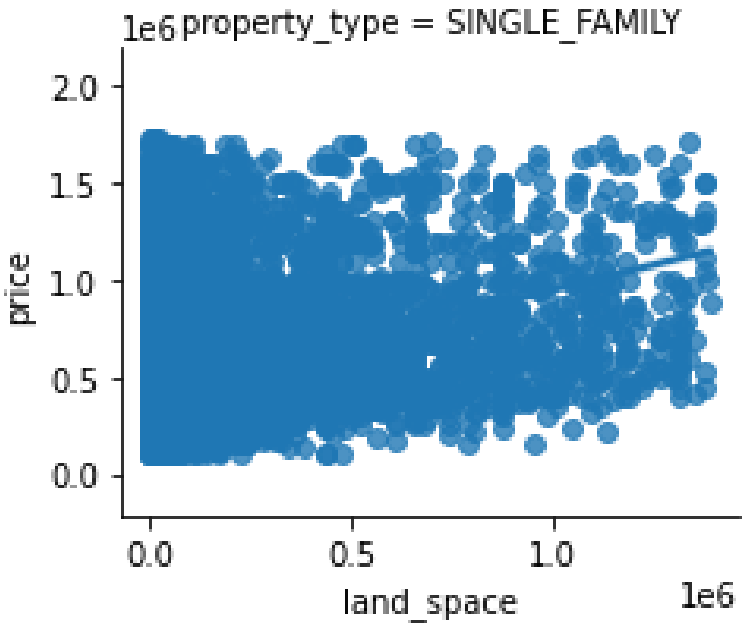
state	count
AK	5011
AR	18347
AZ	37494
BC	59
CA	102464
CO	27927
CT	1
HI	6959
IA	204
ID	12862
IL	43985
KS	11713
LA	24387
ME	1
MO	32039
NE	8434
NJ	1
NM	12703
NV	16585
OK	22472
OR	20557
TX	146636
UT	15268
WA	30460
WY	3430

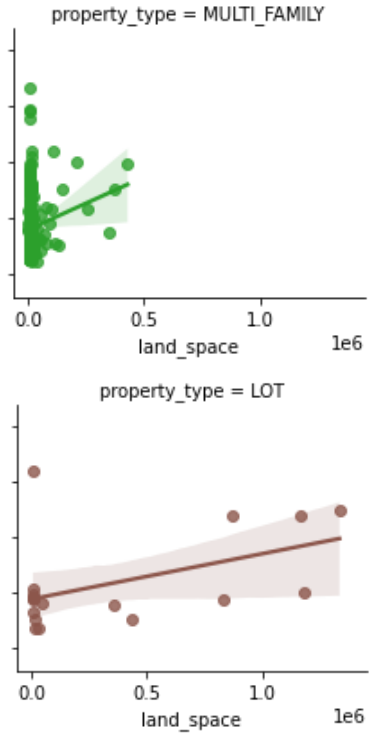


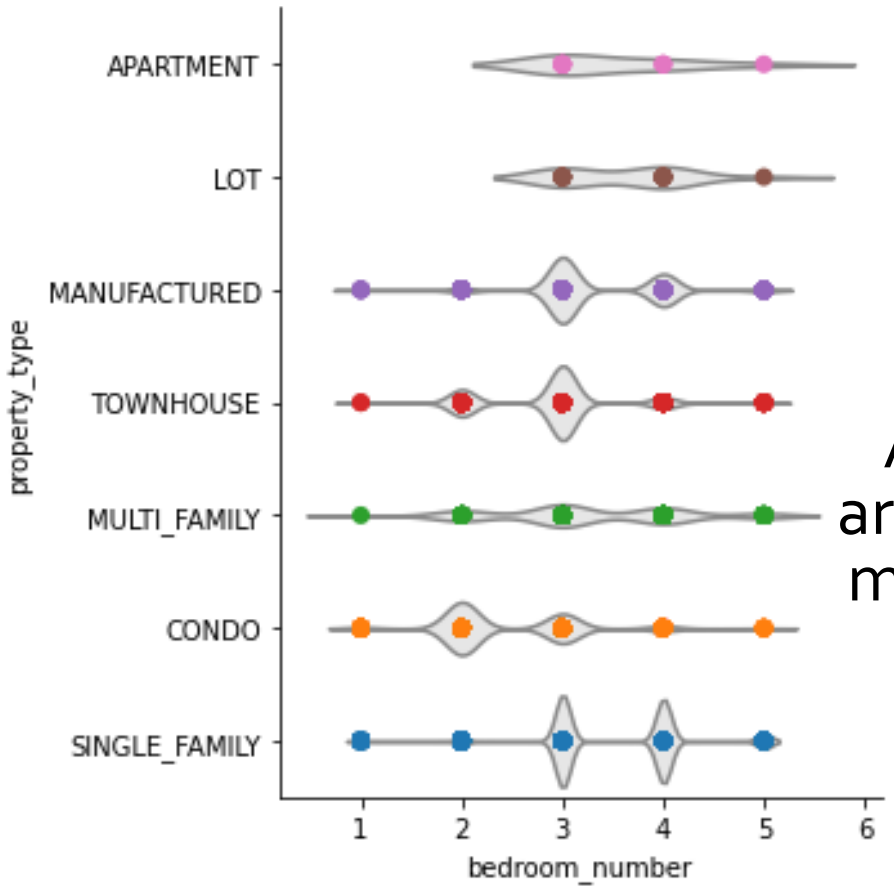
Since this data set contains 600.000 housing data and 28 columns, I realized that the problem could not be solved by the computer after a certain point during feature selection and engineering. In this context, I decided to establish this price estimation system in only one state, and I decided on the state of Texas, which is the state with the most in the data set.

In the graphs below, we can see how the land space and house prices are related according to the property type.

Although there are some exceptions, the conclusion to be drawn from these charts is that there is a proportional increase in the price as the land space increases.







The graph on the left shows the number of bedrooms in each property type.

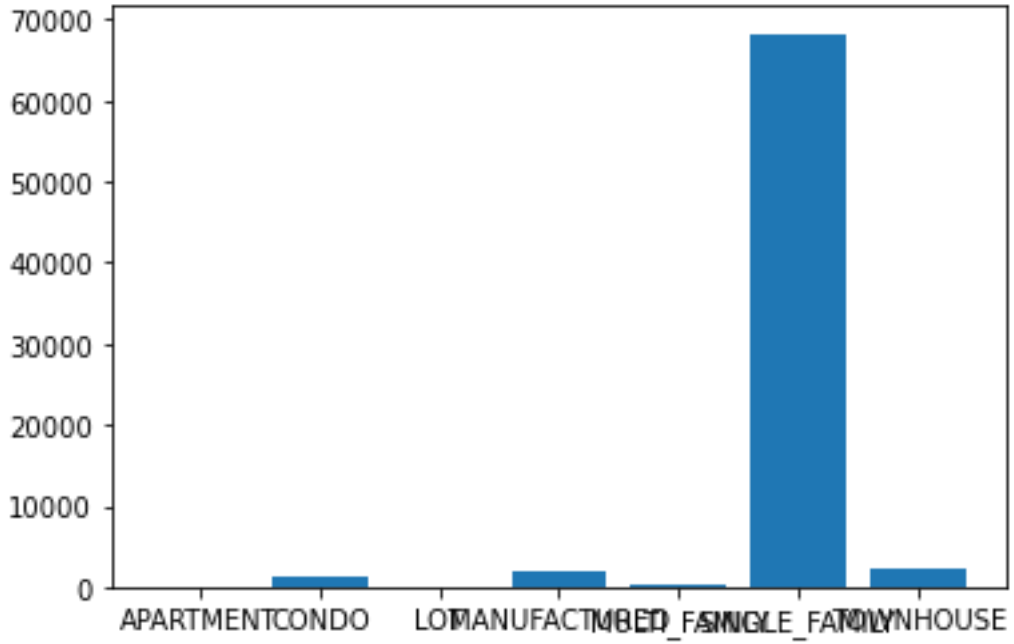
The number of bedrooms and property types are important features in the model we will create.

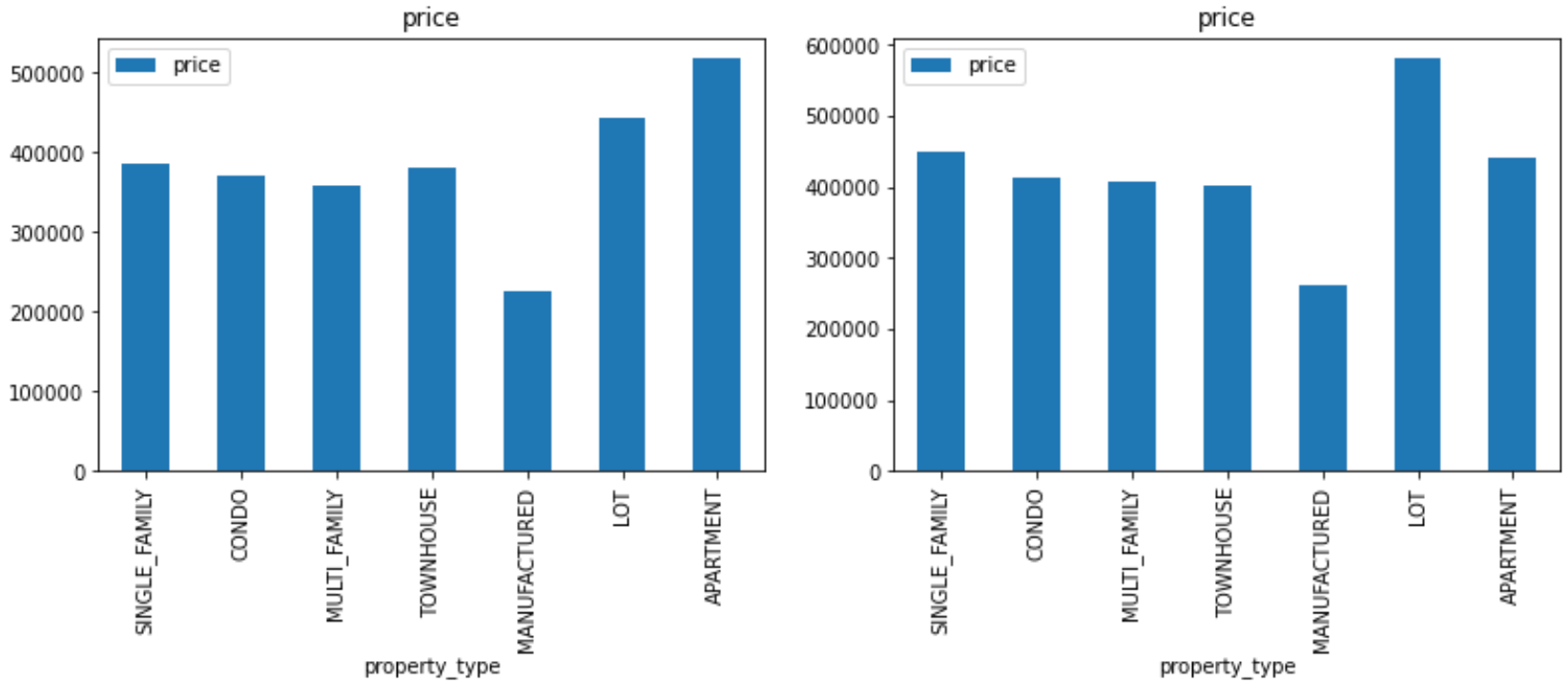
As it can be clearly seen, there are 3-bedroom properties in the manufactured, townhouse, and single-family types, while 2 more bedrooms are in the foreground in condo-type properties.

In the table and graph below, it is seen how many data belong to which property type in the state of Texas in the data set.

As expected, we see the same result in our dataset as there are more single-family homes in the state of Texas, the highlight here is less than 20 data in terms of apartment and lot.

property_type	count
APARTMENT	10
CONDO	1197
LOT	18
MANUFACTURED	1840
MULTI_FAMILY	384
SINGLE_FAMILY	68231
TOWNHOUSE	2162





You can see the median price chart according to the property type in the left chart, and the average price chart by the property type in the right chart.

It is seen that the average price is lower in townhouses and apartments. The opposite is true for lots, and the average lot price is higher than the median lot price, I think the reason for this is the lots with huge spaces.

Modeling

- First Data Cleaning Method contains dropping some rows and features that is not related to price and filling the Na values with the median value.
- Second Data Cleaning Method contains dropping some rows and features but filling the Na values with the zero and expanding dataframe doubling with is_na columns and dummy variables.
- Third Data Cleaning Method doesn't contain dropping some rows and features and fills the Na values with the zero and expanding dataframe doubling with is_na columns and dummy variables.

Model Evaluation Parameters

Parameters represent test scores unless otherwise stated.

		Model	R2 Score	MAE	Test RMSE	Train RMSE
1st Method	0	LinearRegression	0.363	145157.38	267632.05	264324.27
	1	Ridge	0.363	145156.64	267629.87	264322.28
	2	Lasso	-0.068	204574.73	346526.57	348864.72
	3	RandomForestRegressor	0.709	86913.44	181047.34	77939.42
	4	DecisionTreeRegressor	0.473	121493.33	243378.23	1923.78
	5	GradientBoostingRegressor	0.607	107527.61	210256.50	203235.94
	6	LGBMRegressor	0.697	93348.89	184577.41	171888.14
2nd Method	0	LinearRegression	0.942	30045.90	49431.52	48241.73
	1	Ridge	0.942	30046.85	49413.00	48536.19
	2	Lasso	0.450	107510.75	151714.49	151628.97
	3	RandomForestRegressor	0.999	1300.33	4723.14	1532.00
	4	DecisionTreeRegressor	0.998	3063.99	8572.65	14.99
	5	GradientBoostingRegressor	0.998	6016.63	9908.83	9286.04
	6	LGBMRegressor	0.999	3723.60	7822.42	6703.68
3rd Method	0	LinearRegression	0.942	30045.90	49431.52	48241.73
	1	Ridge	0.942	30046.85	49413.00	48536.19
	2	Lasso	0.450	107510.75	151714.49	151628.97
	3	RandomForestRegressor	0.999	1275.67	4720.07	1514.45
	4	DecisionTreeRegressor	0.998	3049.11	8728.56	14.99
	5	GradientBoostingRegressor	0.998	6016.98	9909.13	9286.04
	6	LGBMRegressor	0.999	3723.60	7822.42	6703.68

Modeling

		Model	R2 Score	MAE	Test RMSE	Train RMSE
3rd Method	3	RandomForestRegressor	0.999	1275.67	4720.07	1514.45

The RandomForestRegressor model, modeled using the third method, gives the best results with an RMSE value of 4720, although it is slightly over-fitting since the test RMSE> train is RMSE.

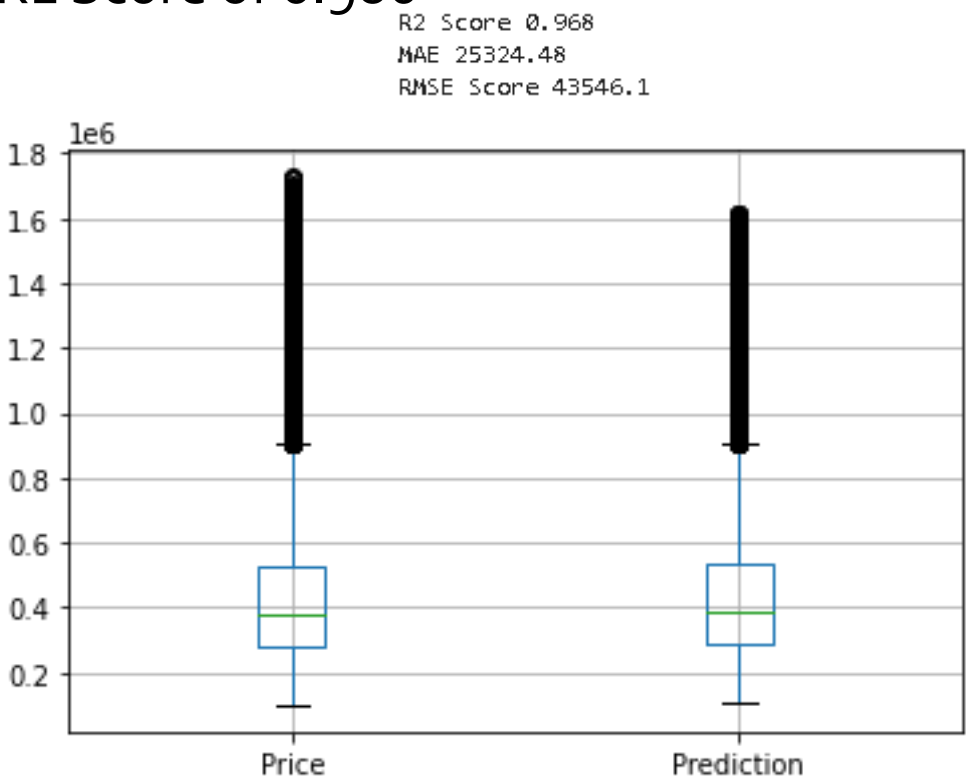
When house prices estimated using the Random Forest regressor method, the result is as follows

There is a slightly higher median value compared to the actual prices. However, the main difference is in the outliers, and some very extreme house prices were found to be much lower prices

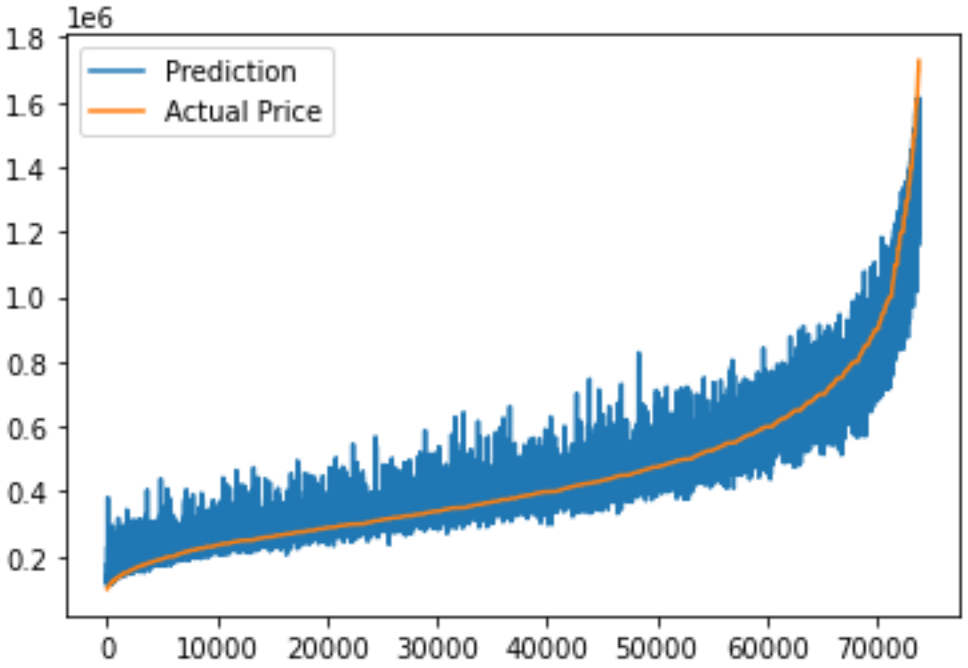
When the model is applied to all data prices predicted with an RMSE value of $43546 and R2 Score of 0.986

	index	Price	Prediction
0	329325	100000.0	125305.94
1	393546	100000.0	124104.20
2	365476	100000.0	131483.99
3	432052	100000.0	143861.90
4	342234	100000.0	173035.80
...	...	...	...
73837	323306	1725000.0	1443478.00
73838	300013	1725000.0	1420492.94
73839	311493	1725000.0	1273033.45
73840	299599	1725000.0	1563419.00
73841	437856	1727225.0	1407071.03

73842 rows × 3 columns



The actual prices vs predictions graph is as follows, as can be seen, the error is more upward-oriented at lower prices, whereas, on the contrary, prices are predicted lower at higher prices.



FUTURE IMPROVEMENTS

- Due to RAM constraints, I had to train just Texas housing data of the original 600000 house dataset. Without resource limitations, I would love to train on the full dataset. Preliminary tests showed that the bigger the training size, the lower the RMSE. One test showed an increase in sample size could increase the RMSE by .03
- With more advanced feature engineering applications, the score of the model can be further increased and the RMSE value and margin of error can be reduced.
- Try to figure out what the attribute "model.coef_" of your trained linear regression model says about which states are the most expensive or how much a square foot of living space costs on average across all states. The weights or slopes are contained in "coef_."