Predicting Home Prices in Los Angeles

Elizabeth Naameh,
VP of Holdings, Redfin

Design

Client: Urban housing markets were particularly disrupted by the Covid pandemic. For investors, this presents an opportunity to capitalize in a new market environment. Redfin wants to understand the LA housing market so that it invest in properties.

Objective: Explore whether the sale price of a home can be modeled against other housing/geographic features.

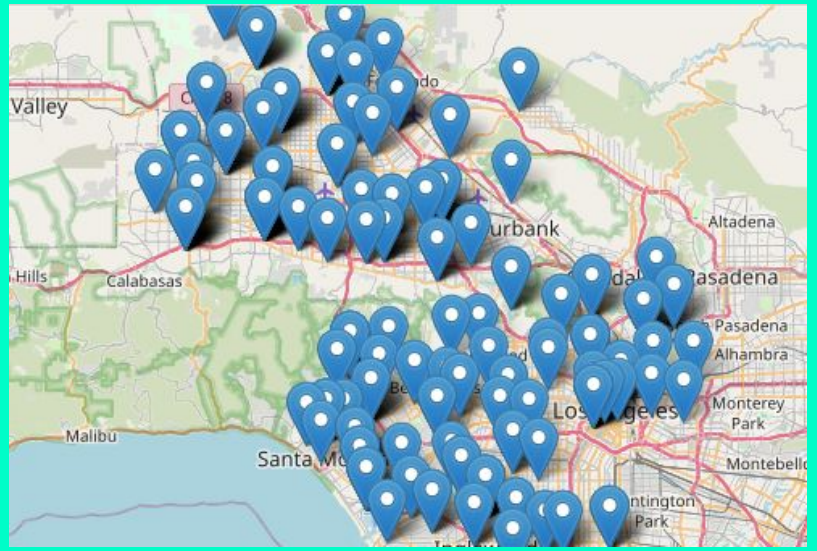
Goal: Produce a regression model that can best interpret a relationship for sale price and a model that can best predict home sale price in Los Angeles.

Data

Data on houses sold in the previous three weeks was scraped from **Zillow** and geographic socio-economic data was scraped from **City-Data.com** by zip code. Each row represents a unique home and address for a ‘Recently Sold’ property in the Los Angeles area.

Of 600+ home sales scraped, 288 were used in our final analysis.

- BeautifulSoup & Selenium
- Numpy & Pandas
- Scikit-learn & Statsmodels
- Matplotlib & Seaborn
- Folium, Geopandas & Geopy

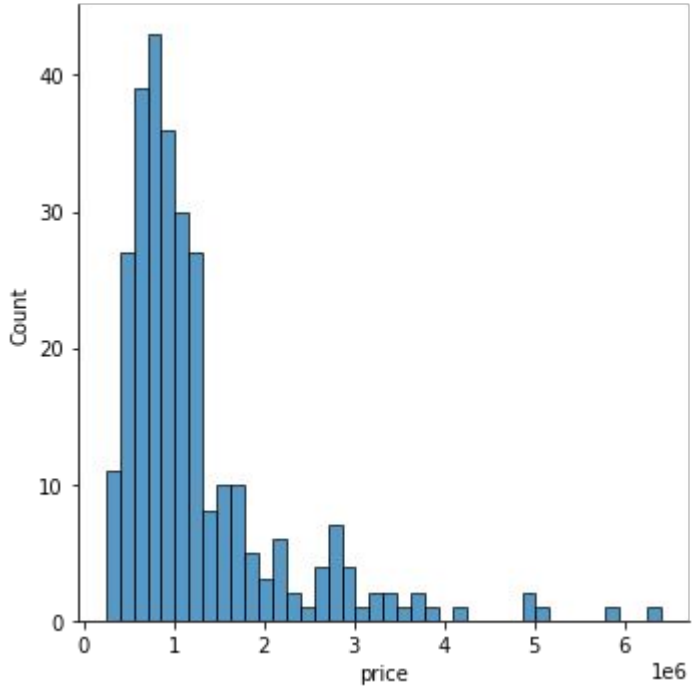


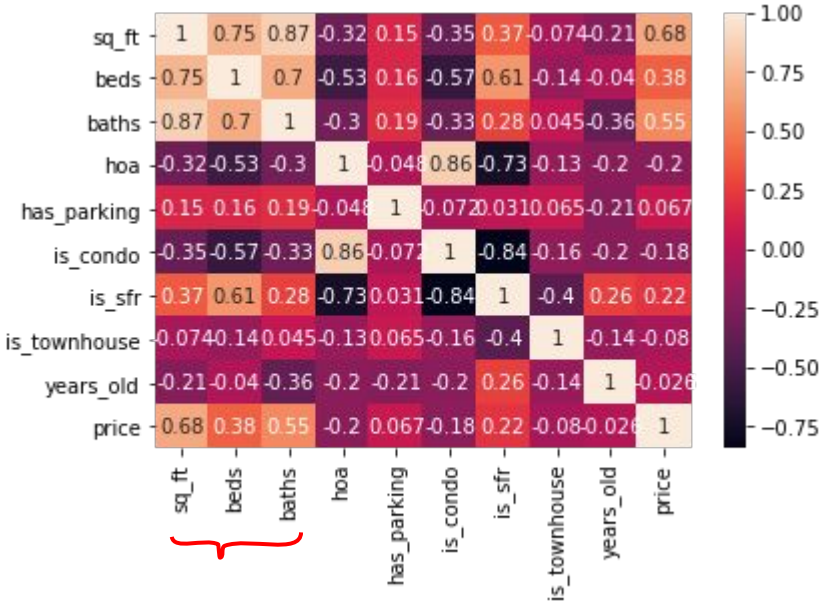
Data Cleaning & EDA

- Delete duplicates and drop rows with null *'prices', 'addresses'*
- Binarize categorical data
  - New columns include *'has_parking', 'is_condo', 'is_sfr', 'is_townhouse'*
- Restrict sold listings to prior three weeks
- Impute missing *'hoa' (0)* and *'beds'* (0, indicates studio)

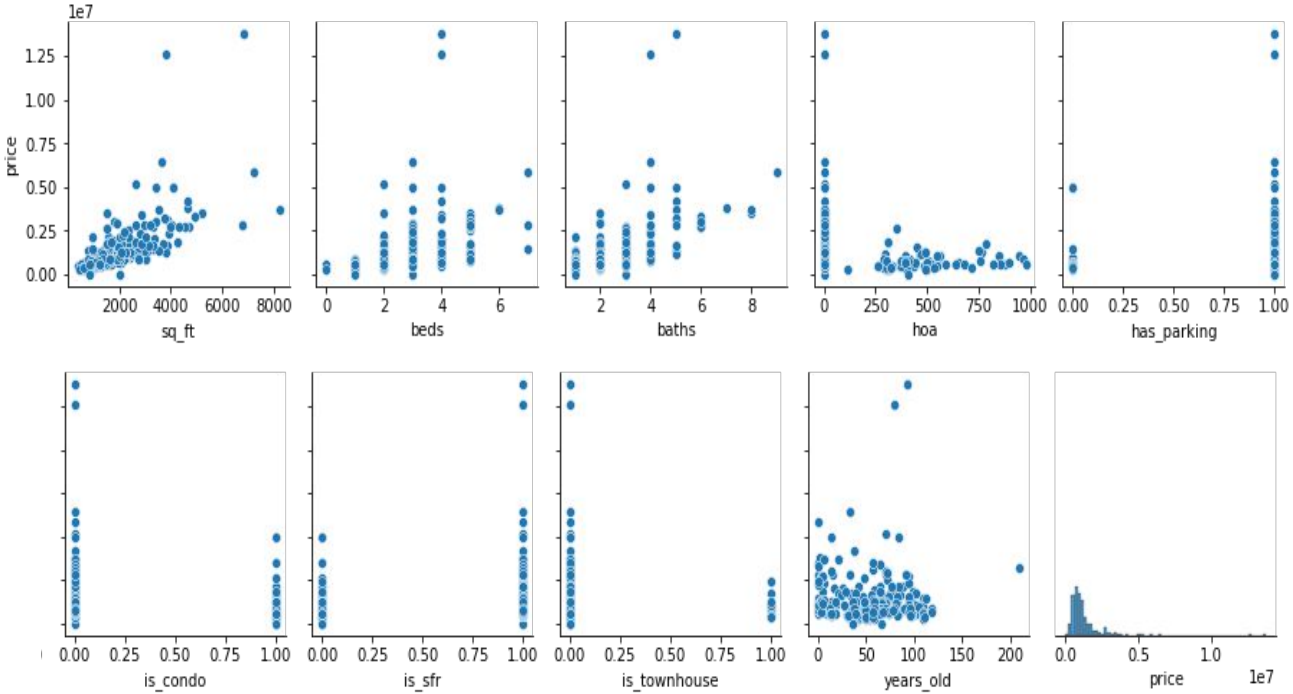
	address	price	sq_ft	date_sold	beds	baths	year_built	hoa	parking	type_
0	1135 W 185th St, Gardena, CA 90248	800000.0	1866.0	07/07/21	3.0	3.0	1954.0	NaN	2 Attached Garage spaces	Single Family Residence
1	9631 Compton Ave, Los Angeles, CA 90002	610000.0	1522.0	07/07/21	4.0	3.0	2021.0	NaN	2 Attached Garage spaces	Single Family Residence
2	8701 Delgany Ave UNIT 304, Los Angeles, CA 90293	1125000.0	1785.0	07/07/21	3.0	3.0	1964.0	460.0	2 Garage spaces	Condominium
3	906 Parkman Ave, Los Angeles, CA 90026	1455000.0	NaN	07/07/21	NaN	NaN	1937.0	NaN	Carport	Multi Family

Baseline Model





Baseline Model



OLS
Train: 0.579
Val: 0.235

Import Geographic Data

- Scrape socio-economic data by Zip Code
  - *'median_resident_age'*
  - *'avg_household_size'*
  - *'avg_household_income'*
  - *'pct_poverty'*
  - *'pct_bachelors'*
- Merge datasets on zip_code

	zip_code	median_age	avg_household	median_income	pct_poverty	pct_bachelors	url
0	90248	43.5	3.0	64253.0	9.7	27.9	http://www.city-data.com/zips/90248.html
1	90002	28.3	4.1	38987.0	28.8	6.1	http://www.city-data.com/zips/90002.html
2	90293	38.3	1.9	110698.0	5.8	74.7	http://www.city-data.com/zips/90293.html

Feature Selection

Drop features with high multicollinearity and low correlation.

- *'hoa'*
- *'median_age'*
- *'has_parking'*
- *'is_townhouse'*

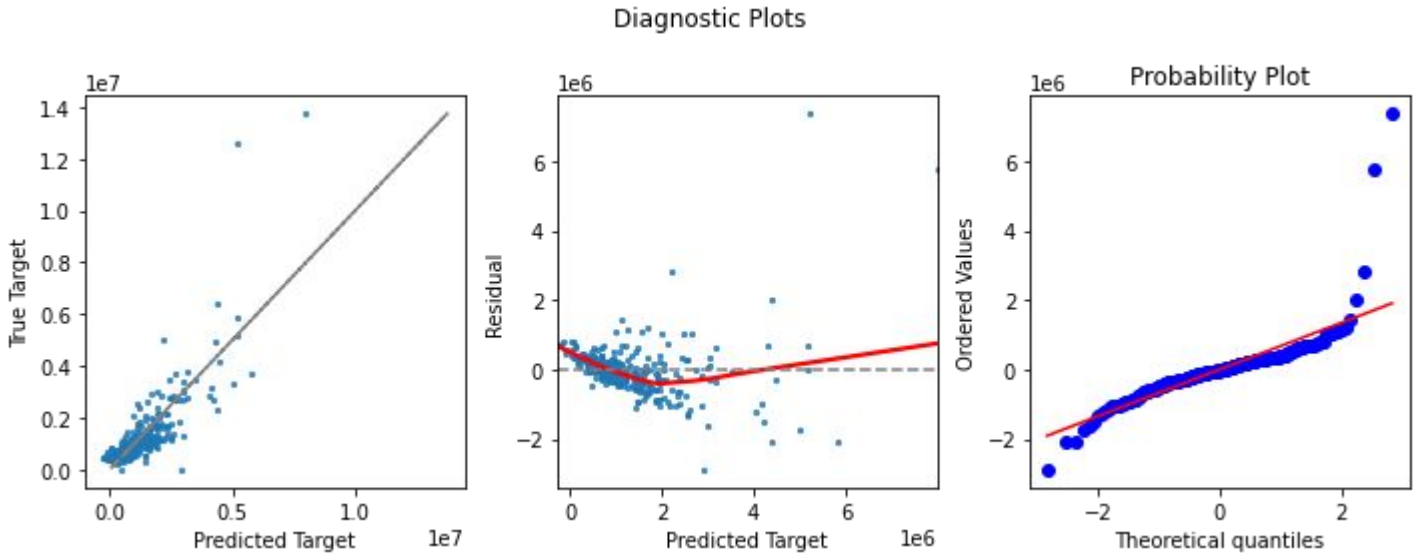
Improvement on baseline:

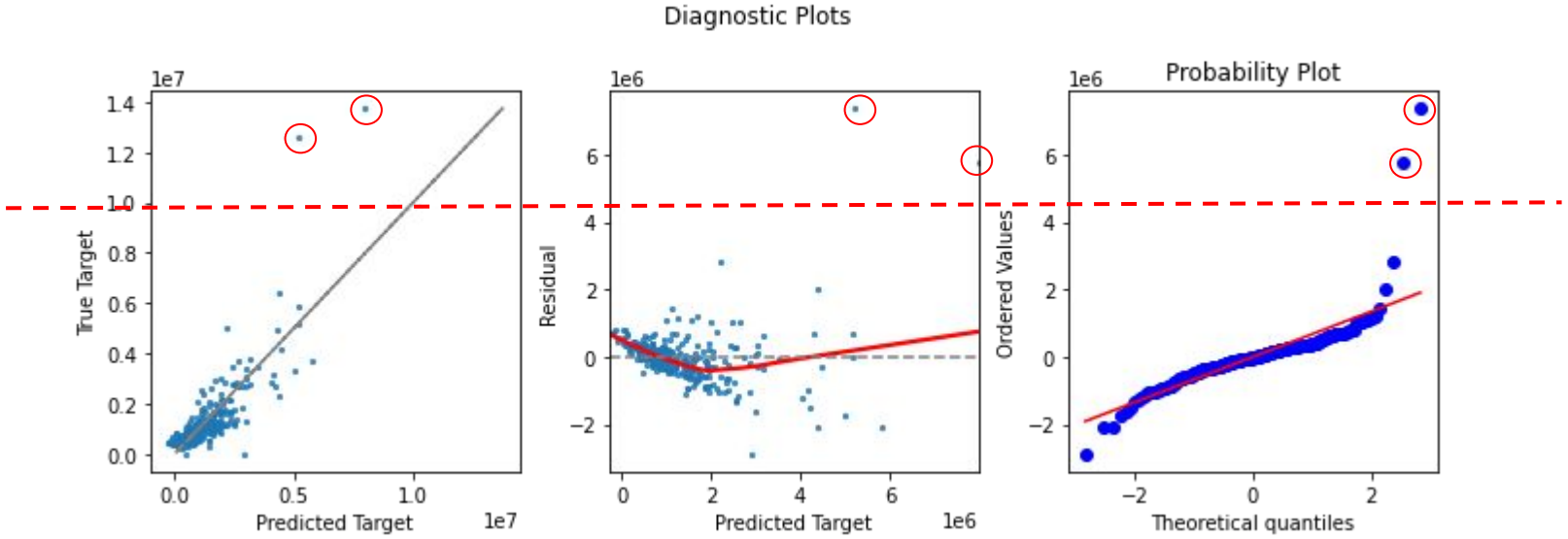
```
OLS
Train:  0.693
Val:  0.356
```

Variance Inflation Factor

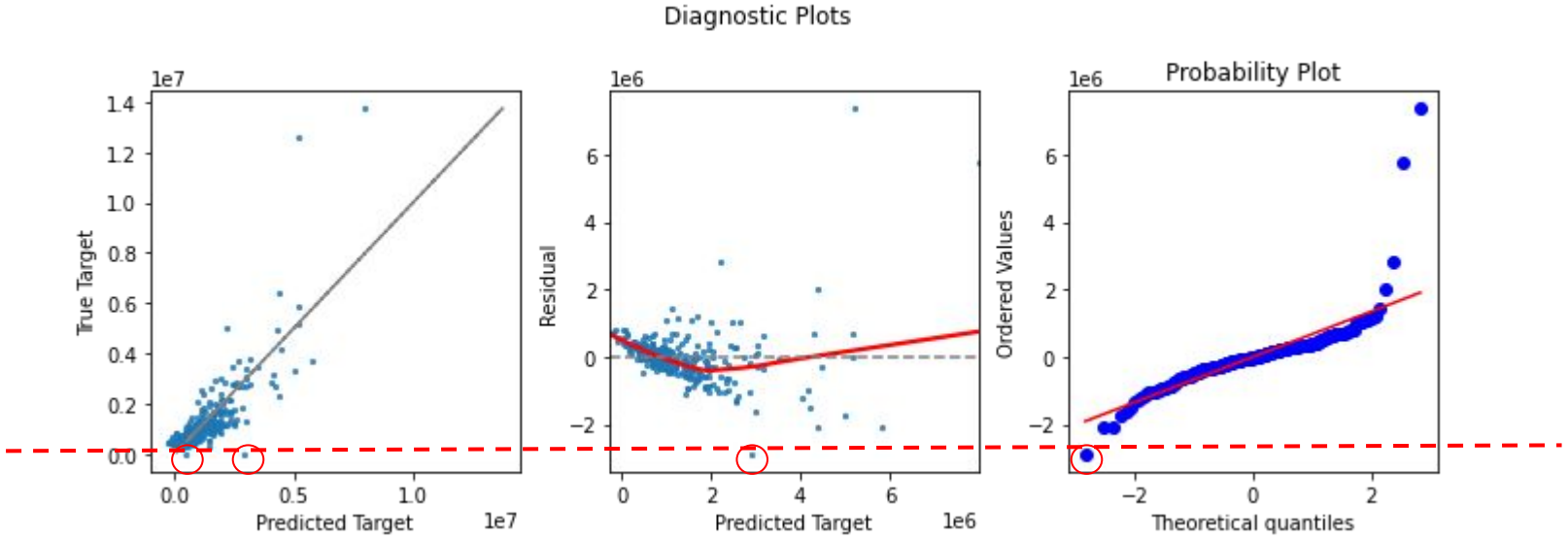
sq_ft	20.850978
beds	28.544846
baths	27.677189
is_condo	4.953304
is_sfr	13.496346
years_old	5.686895
avg_household	37.070400
median_income	47.674372
pct_poverty	10.101321
pct_bachelors	34.739924

dtype: float64


Diagnostic Plots
1e7
True Target
Predicted Target
1e7
1e6
Residual
Predicted Target
1e6
1e6
Ordered Values
Probability Plot
Theoretical quantiles



	address	y_pred	y_actual	residual
181	1650 Amalfi Dr, Pacific Palisades, CA 90272, USA	7.983059e+06	13750000.0	5.766941e+06
157	992 Napoli Dr, Pacific Palisades, CA 90272, USA	5.225908e+06	12600000.0	7.374092e+06
243	1002 Alta Ave, Santa Monica, CA 90402, USA	4.368953e+06	6402725.0	2.033772e+06



	address	y_pred	y_actual	residual
287	222 S Central Ave APT 238, Los Angeles, CA 900...	4.794929e+05	2000.0	-4.774929e+05
91	2201 Coldwater Canyon Dr, Beverly Hills, CA 90...	2.890890e+06	7100.0	-2.883790e+06
201	5460 White Oak Ave UNIT A227, Encino, CA 91316...	3.074047e+05	242000.0	-6.540468e+04

Zillow

Save Share More

4 bd | 4 ba | 3,802 sqft

992 Napoli Dr, Pacific Palisades, CA 90272

Sold: $12,600,000 | Sold on 06/30/21 | Zestimate®: **$12,666,900**

Est. refi payment: $54,449/mo **Refinance your loan**

Home value | Owner tools | Home details | Neighborhood details | Similar homes

Zillow

Save Share More

4 bd | 5 ba | 6,800 sqft

1650 Amalfi Dr, Pacific Palisades, CA 90272

Sold: $13,750,000 | Sold on 06/29/21 | Zestimate®: **$13,832,200**

Est. refi payment: $59,391/mo **Refinance your loan**

Home value | Owner tools | Home details | Neighborhood details | Similar homes

Feature Selection

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.731
Model:	OLS	**Adj. R-squared:**	0.722
Method:	Least Squares	**F-statistic:**	75.43
Date:	Fri, 09 Jul 2021	**Prob (F-statistic):**	4.04e-73
Time:	07:28:40	**Log-Likelihood:**	-4177.3
No. Observations:	288	**AIC:**	8377.
Df Residuals:	277	**BIC:**	8417.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.832e+06	4.87e+05	-3.761	0.000	-2.79e+06	-8.73e+05
sq_ft	493.6607	64.750	7.624	0.000	366.196	621.126
beds	-1.719e+05	4.93e+04	-3.490	0.001	-2.69e+05	-7.49e+04
baths	1.47e+05	5.29e+04	2.778	0.006	4.28e+04	2.51e+05
is_condo	6.84e+04	1.31e+05	0.523	0.601	-1.89e+05	3.26e+05
is_sfr	5.772e+05	1.29e+05	4.459	0.000	3.22e+05	8.32e+05
years_old	2544.0683	1155.414	2.202	0.028	269.561	4818.576
avg_household	-2.334e+04	1.23e+05	-0.190	0.849	-2.65e+05	2.18e+05
median_income	9.0275	2.437	3.705	0.000	4.230	13.825
pct_poverty	3.613e+04	7489.767	4.824	0.000	2.14e+04	5.09e+04
pct_bachelors	1.338e+04	5314.985	2.518	0.012	2920.668	2.38e+04

Drop non-significant features

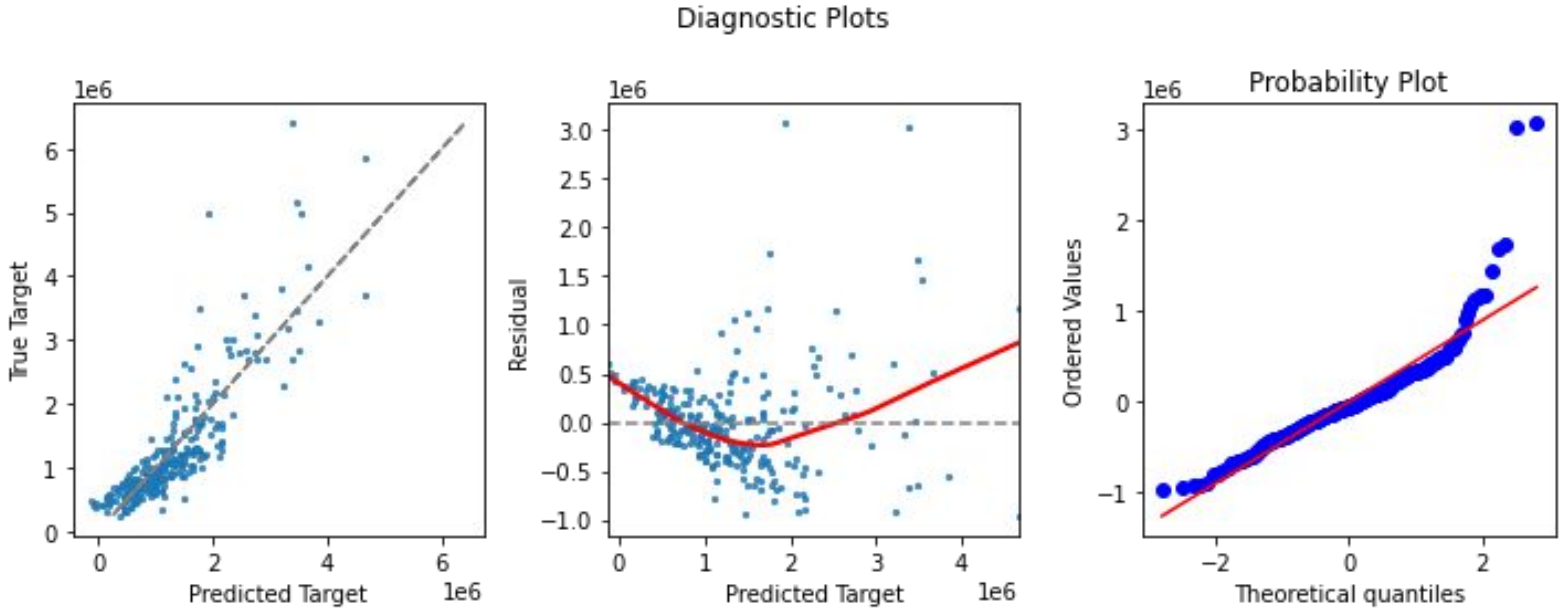
- *'avg_household', 'is_condo'*

Improving on Baseline Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.731
Model:	OLS	**Adj. R-squared:**	0.723
Method:	Least Squares	**F-statistic:**	94.83
Date:	Fri, 09 Jul 2021	**Prob (F-statistic):**	4.94e-75
Time:	07:32:57	**Log-Likelihood:**	-4177.4
No. Observations:	288	**AIC:**	8373.
Df Residuals:	279	**BIC:**	8406.
Df Model:	8		
Covariance Type:	nonrobust		

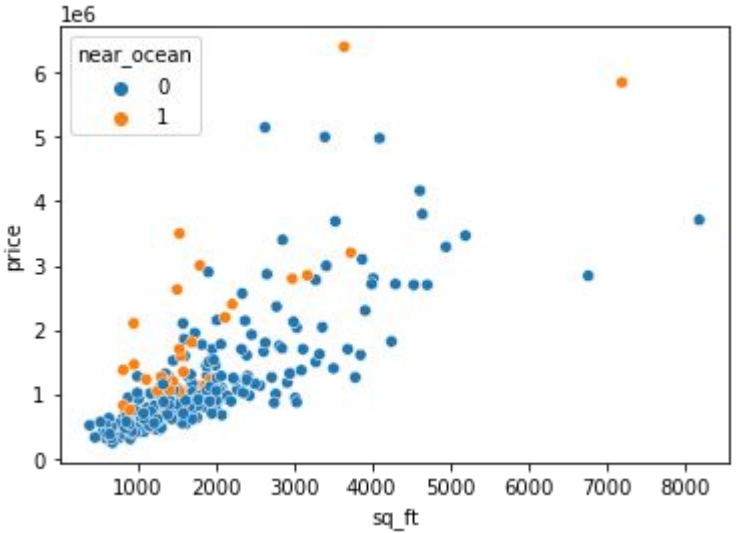
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.865e+06	2.47e+05	-7.550	0.000	-2.35e+06	-1.38e+06
sq_ft	500.4162	62.459	8.012	0.000	377.465	623.367
beds	-1.742e+05	4.76e+04	-3.661	0.000	-2.68e+05	-8.06e+04
baths	1.403e+05	5.13e+04	2.735	0.007	3.93e+04	2.41e+05
is_sfr	5.299e+05	9.4e+04	5.640	0.000	3.45e+05	7.15e+05
years_old	2525.3727	1122.636	2.250	0.025	315.460	4735.286
median_income	8.7311	1.965	4.444	0.000	4.864	12.599
pct_poverty	3.677e+04	7373.027	4.988	0.000	2.23e+04	5.13e+04
pct_bachelors	1.449e+04	2743.429	5.280	0.000	9084.802	1.99e+04


Diagnostic Plots
1e6
True Target
Predicted Target
1e6
1e6
Residual
Predicted Target
1e6
1e6
Probability Plot
Ordered Values
Theoretical quantiles

Feature Engineering

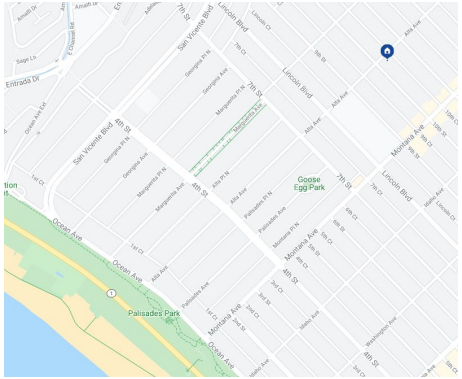
We continue to underperform for unique properties,
e.g. sweeping downtown views, beachfront locations.

Create new variable, `*near_ocean*`

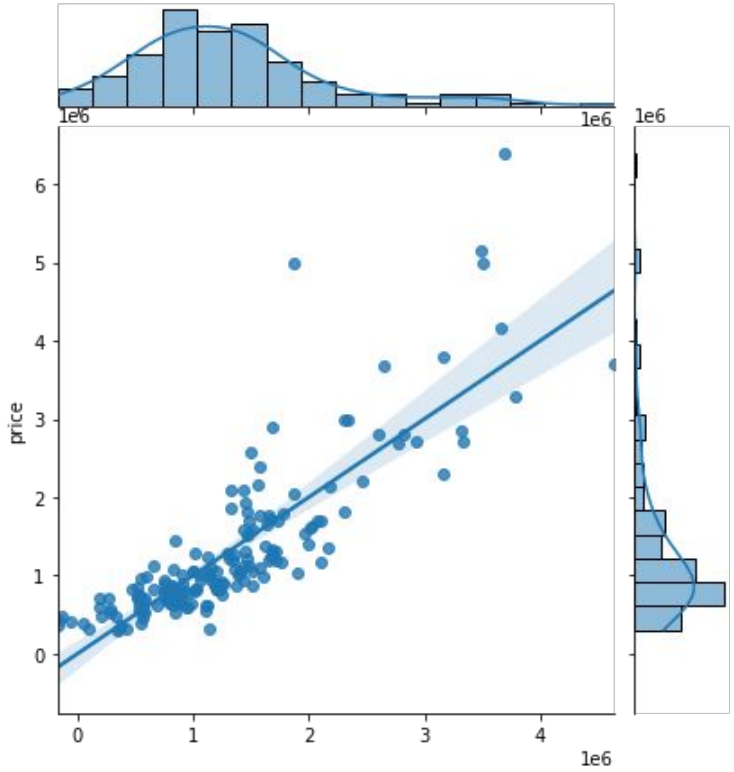



```
OLS
Train:  0.709
Val:  0.857
```

```
RMSE:  530052.544
MAE:  355690.896
```



Final Interpretive Model



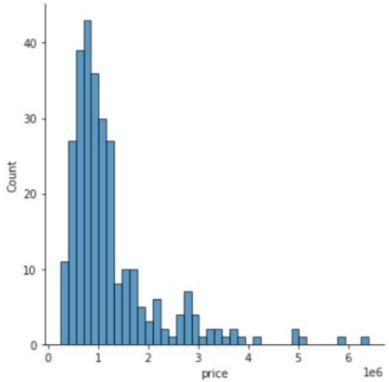

```
Repeated Cross Validation Results:

Simple mean cv r^2: -0.776 +- 0.033
Ridge mean cv r^2: -0.776 +- 0.033
```

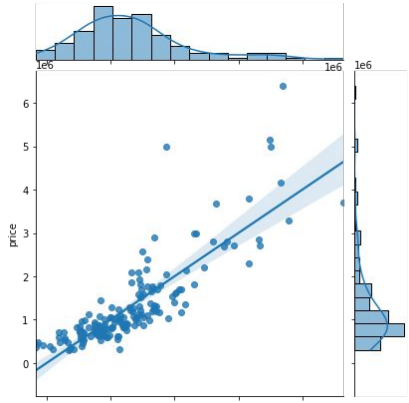
	coef	P>|t|
const	-1.738e+06	0.000
sq_ft	528.1412	0.000
beds	-1.771e+05	0.000
baths	1.205e+05	0.019
is_sfr	5.685e+05	0.000
years_old	2082.2522	0.063
median_income	7.7889	0.000
pct_poverty	3.491e+04	0.000
pct_bachelors	1.32e+04	0.000
near_ocean	2.955e+05	0.003

Final Predictive Model

Before log transformation

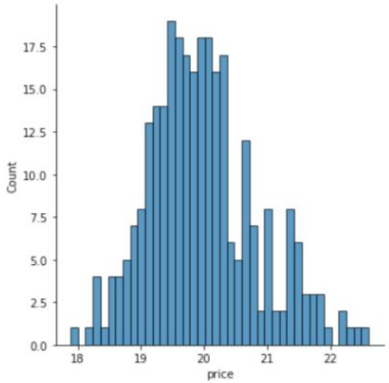
Target = prices

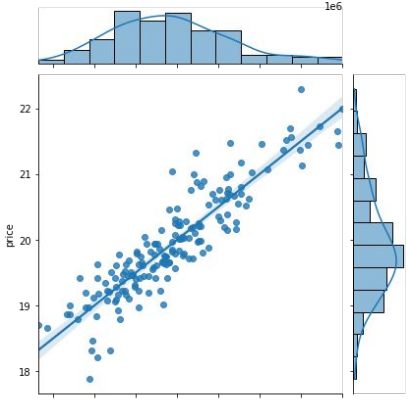




After log transformation

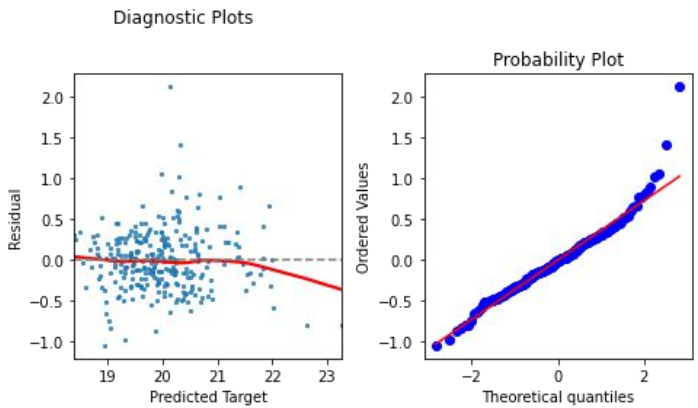
Target = log2(prices)






```
Linear Regression val R^2: 0.828
Ridge Regression val R^2: 0.829
Degree 2 polynomial regression val R^2: 0.697
```

```
OLS with log2(price)
Train:  0.815
Val:  0.828
RMSE:  0.344
MAE:  0.262
```



Future Work

To improve our model, include additional features:

- Lot size
- Has view or Floor Number (if condo)
- Proximity to ocean (miles)

To improve business relevance:

- Predict rate of return for investment properties
- Model sale price vs. predicted monthly rent