

Business Case


Real estate is the single largest asset class. Having models for predicting fair market values for real estate properties would be of interest to buyers, sellers, lenders, investors, and municipalities as well as other market participants. It is likely that different participants will have different requirements regarding model performance. Acceptable level of model performance is likely to be $R^2 > 0.9$. The data for condo listings in Miami was obtained from Redfin. A total of 350 listings were downloaded.

Q1. Create a feature list to include the following features: Bedrooms, Bathrooms, Square Footage, Lot Size, Property type.

Project dataset:
miami_beach_condos.csv

Features:
28

Datapoints:
351

 Data Quality Assessme...
For BedsBathsSqFt

2

View info

Menu




Search

Feature List: BedsBathsSqFt

View Raw Data

Create feature list

1-4 of 4

<input type="checkbox"/> Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
<input type="checkbox"/> PRICE		8	Target	Numeric	204	0	1,385,485	2,164,423	550,000	89,000	1.81e+7
<input type="checkbox"/> SQUARE FEET		12	<div></div>	Numeric	205	3	1,106	714	910	112	4,904
<input type="checkbox"/> BATHS		10	<div></div>	Numeric	9	3	1.74	0.84	1.50	1	6.50
<input type="checkbox"/> BEDS		9	<div></div>	Numeric	6	0	1.47	0.90	1	0	5

Q2. Build a linear regression model using DataRobot and report the following metrics for cross-validation and holdout samples: R2, MAPE, MAE, RMSE. (12 pts.)

Linear Regression

Missing Values Imputed | Standardize | Linear Regression

BedsBathsSqFt

64.0 %

0.4586

0.6739

0.6874

M4 BP40 REF β_i

R ²	Cross Validation	Holdout
	0.67	0.69

Linear Regression

Missing Values Imputed | Standardize | Linear Regression

BedsBathsSqFt

64.0 %

6.8630e+5

6.9743e+5

1.2387e+6

M4 BP40 REF β_i

MAE	Cross Validation	Holdout
	\$6.974 million	\$1.2387 million

<div> <div>Linear Regression</div> <div>Missing Values Imputed Standardize Linear Regression</div> <div>M4 BP40 REF β_i</div> </div> <div> <div>BedsBathsSqFt</div> <div>64.0 %</div> <div>1.1094e+6</div> <div>1.1410e+6</div> <div>3.5005e+6</div> </div>		
RMSE	Cross Validation	Holdout
	\$1.141 Million	\$3.5 million

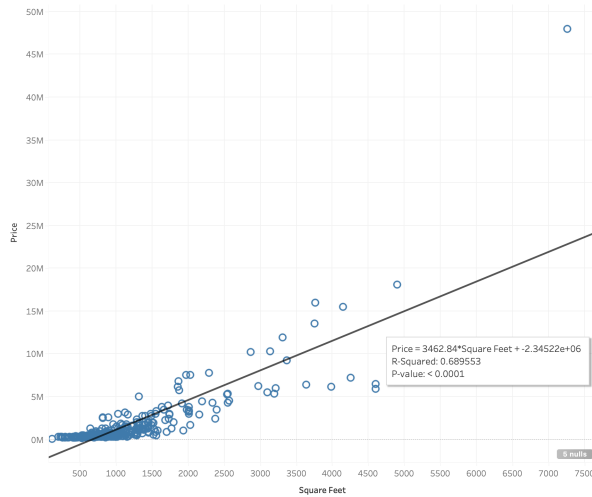
<div> <div>Linear Regression</div> <div>Missing Values Imputed Standardize Linear Regression</div> <div>M4 BP40 REF β_i</div> </div> <div> <div>BedsBathsSqFt</div> <div>64.0 %</div> <div>95.9708</div> <div>94.7816</div> <div>94.9452</div> </div>		
MAPE	Cross Validation	Holdout
	94.8%	94.9%

R^2 suggests an unacceptable level of performance, given the target value for R^2 is greater than 0.9 and we saw values of 0.67 and 0.69 for the cross validation and holdout, respectively. However, the other metrics also indicate issues with the model:

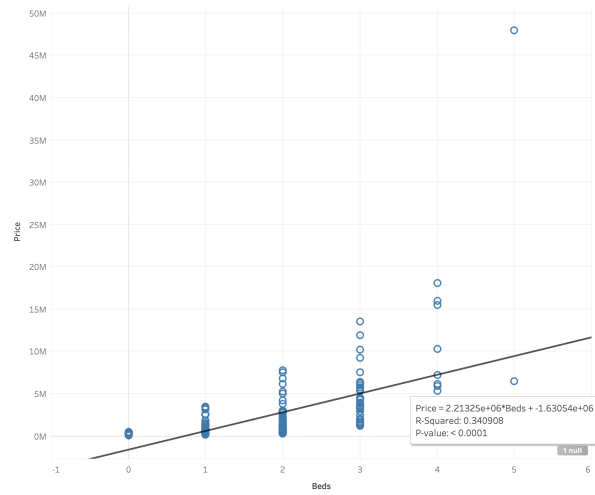
- The MAE is quite high, especially for the cross-validation sample at approximately \$6.974 million, and lower for the holdout sample at approximately \$1.2387 million. This indicates the average magnitude of the errors in the predictions is substantial, particularly for the cross-validation sample.
- The RMSE for the cross-validation sample is \$1.141 million, which is lower than the MAE, suggesting that there are not many large errors inflating the RMSE; however, the RMSE for the holdout is much higher at \$3.5 million, which is a concern.
- The MAPE is extremely high for both cross-validation and holdout, at 94.8% and 94.9%, respectively. This indicates that the average percentage error between the model's predictions and the actual values is very high, which is problematic because it suggests the model's predictions are, on average, off by a very large percentage.

Q3. Visualize the effects of square footage, number of bedrooms and number of bathrooms on the property value using Tableau. Use trend analysis to estimate R^2 reflecting the effects of individual property features on property values. Report R^2 for beds, baths, sqft or lot size in relation to the asking price. (5 pts.)

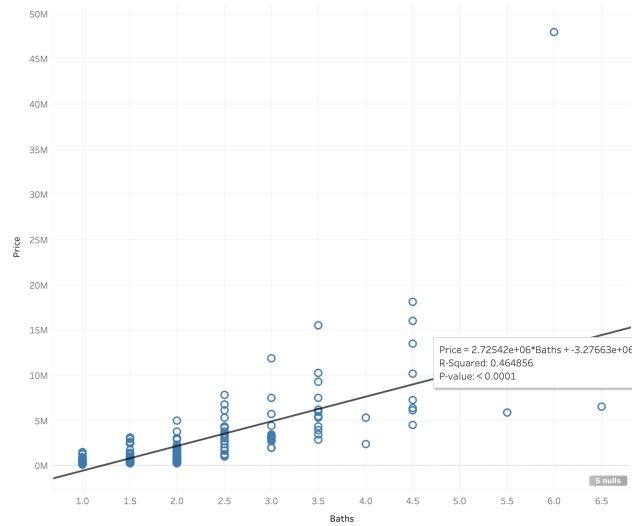
Price vs. Square Footage



Price vs. Bedrooms



Price vs. Baths



R² values for simple linear regressions:

Square Feet	0.69
Beds	0.34
Baths	0.46

Q4. What is the single best predictor of real estate prices in your data based on R²? Does it make sense? (3 pts.)

The R² value, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variable. It ranges from 0 to 1, where a value closer to 1 indicates a better fit and suggests that the model explains a larger portion of the variance in the dependent variable. In the data, square footage has the highest R² value of 0.69, indicating that it is the best predictor of real estate prices among the variables we considered (square footage, number of

bedrooms, and number of bathrooms). This suggests that square footage explains a significant portion of the variance in asking prices for condos in the Miami Beach market. This finding makes intuitive sense because the size of a property is often a primary factor that potential buyers consider when assessing its value. Larger properties typically offer more space and utility, which can directly influence their market price. While the number of bedrooms and bathrooms also affects property values, these features may have a more indirect impact, reflected in their lower R^2 values. Bedrooms and bathrooms contribute to the overall utility and appeal of a property but may not capture the price variance as effectively as the total square footage, which provides a more comprehensive measure of a property's size and potential value.