Case Study – Real Estate Prices

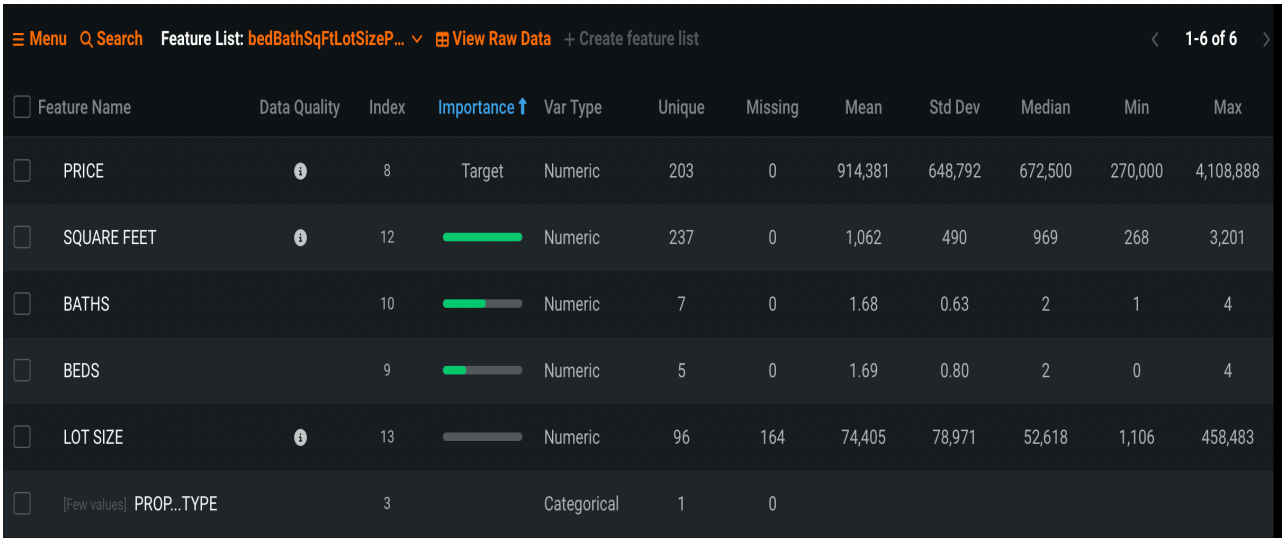
Business Case:
Real estate is a huge market and having good models that can make predictions on the price of the property can be of great help when making important decisions such as the purchase of a home or real estate as an investment vehicle. It can also be an important tool for a real estate sales agency, since it will allow them to estimate the sale value of the real estate which for them in this case are assets. Acceptable level of model performance is likely to be R2 >= 0.9
Dataset is of San Diego condos and was obtained from Redfin. A total of 350 records were downloaded.

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

   Data Pre-processing –

   The data originally has 28 features, but we are only using the features listed below -

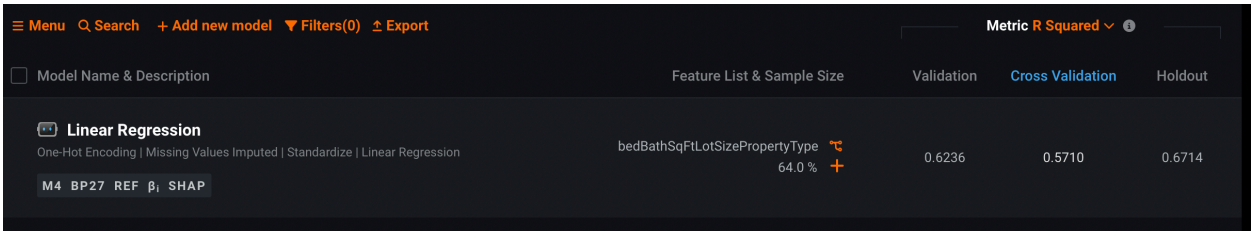
   Number of Bedrooms, Number of Bathrooms, Square Footage, Lot Size and Property type. We have only downloaded the data for condos hence the property type feature has only 1 unique value.



Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
PRICE		8	Target	Numeric	203	0	914,381	648,792	672,500	270,000	4,108,888
SQUARE FEET		12		Numeric	237	0	1,062	490	969	268	3,201
BATHS		10		Numeric	7	0	1.68	0.63	2	1	4
BEDS		9		Numeric	5	0	1.69	0.80	2	0	4
LOT SIZE		13		Numeric	96	164	74,405	78,971	52,618	1,106	458,483
[Few values] PROP...TYPE		3		Categorical	1	0					

2. **Build a linear regression model using DataRobot and report the following metrics for cross-validation and holdout samples: R2, MAPE, MAE, RMSE.**

Model – Multiple Linear Regression model



Metrics	Cross-validation	Holdout
R2	0.57	0.67
MAPE	29.12%	30.50%
MAE	$0.27 million	$0.32 million
RMSE	$0.41 million	$0.57 million

The value of R2 on the cross-validation set is too low suggesting that the model doesn't fit well. The RMSE of $0.41 million for property prices suggests that on average, the predictions made by the model have an error of $0.41 million when compared to the actual observed property prices which is not quite acceptable.
The metrics also do not seem to be too promising for the unseen holdout data, giving an R2 of 0.67 and MAE and RMSE of $0.32 million and $0.57 million respectively.

3. **Visualize the effects of square footage, number of bedrooms and number of bathrooms on the property value using Tableau. Use trend analysis to estimate R2 reflecting the effects of individual property features on property values. Report R2 for beds, baths, sqft or lot size in relation to the asking price.**

Simple Linear Regression -

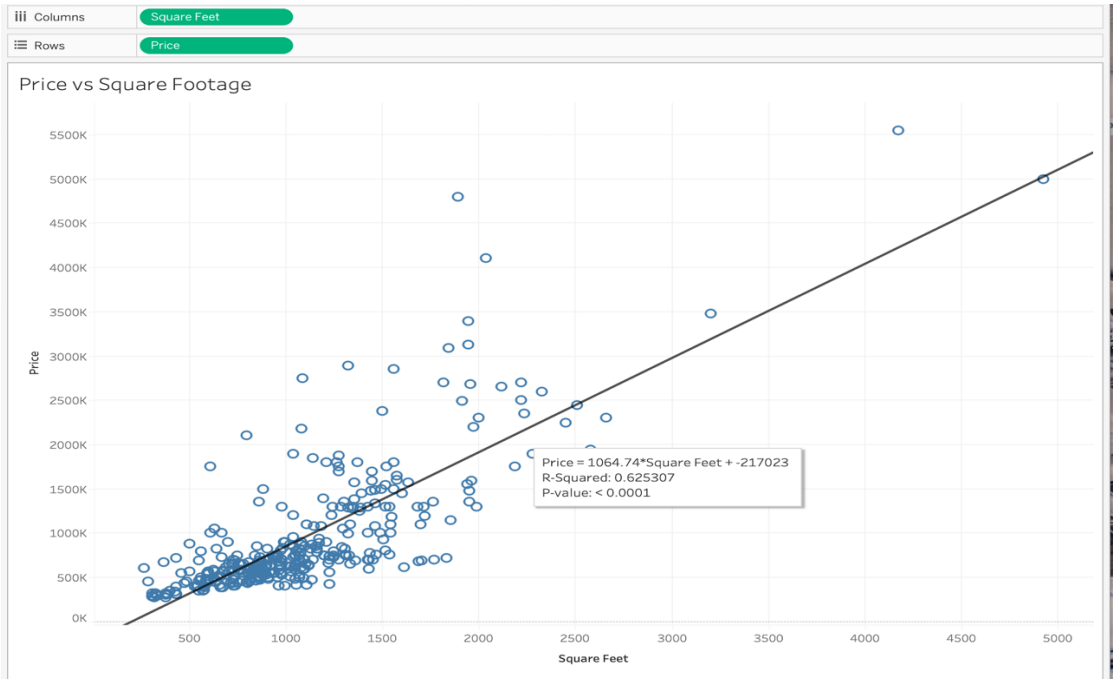


Fig: Scatter plot for Square Ft vs Price and its regression line

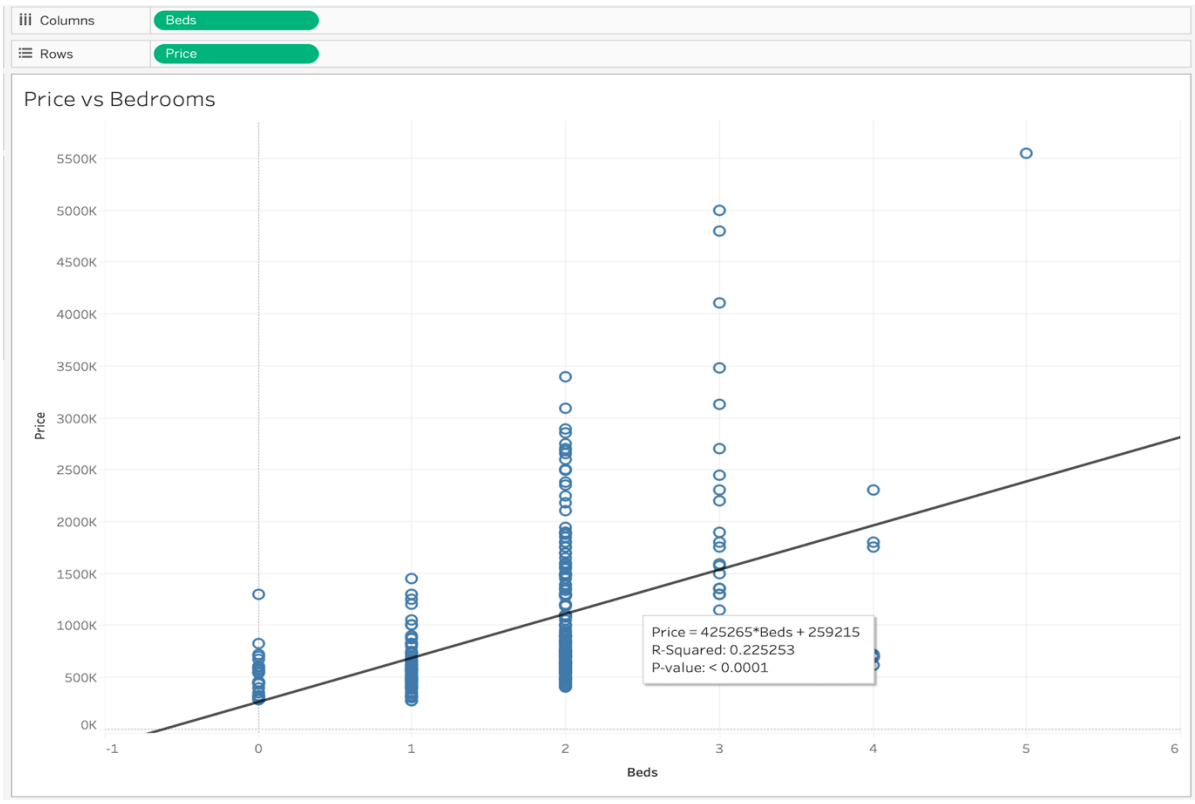


Fig: Scatter plot for Number of bedrooms vs Price and its regression line

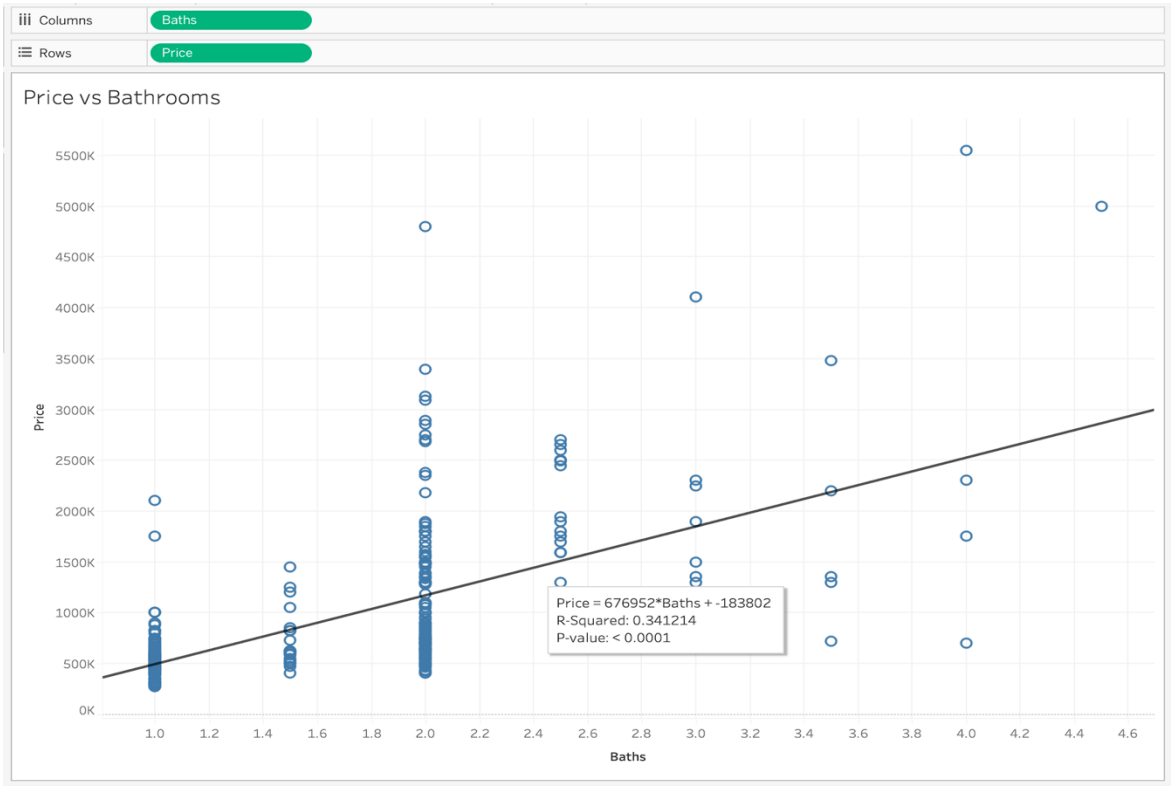


Fig: Scatter plot for Number of bathrooms vs Price and its regression line

R2 values for the simple linear regression are as follows –

Square Ft	0.62
Beds	0.22
Baths	0.34

The above R2 values suggets that square ft has the strongest relationship with the value to be predicted that is the price of the property, followed by a weaker association of number of bathrooms while the number of bedrooms has the weakest prediction power for the price.

4. **What is the single best predictor of real estate prices in your data based on R2? Does it make sense?**

Based on R2 values, square footage is the single best predictor of asking prices in the San Diego condo market. It does make sense as we expect the number of baths, the number of beds and square footage of the property to be correlated, since adding a bedroom or a bathroom to the property means adding to its square footage also.
It can also be visualized by the below bar graph which shows the percentage effect that the different features have over our prediction feature that is price.