Gregory Buisson 02/03/2024

Analysis of Multiple Regression Modeling for House Prices

Introduction and Business Understanding:

The study's objective is to create a robust regression model to forecast house prices in Dallas, TX, where the real estate market's dynamic pricing presents a prime opportunity for applying machine learning to enhance prediction precision. Key variables have been selected to accurately predict housing prices. The approach includes an explanatory model to assess the significance of various predictors on price and a predictive model to estimate housing prices based on these predictors' interactions with the target variable.

Data Understanding:

Before launching the regression analysis, the study required a meticulous preparation of existing Redfin data. This step ensured the data's integrity, essential for reliable modeling and analysis. The process included cleaning and structuring the data, with a specific focus on location information, which was streamlined through binning. Utilizing the tabulate function, two tables were crafted: one displaying the average property price by ZIP code, treated as a nominal variable, and another showcasing the average price by location. Both datasets were rigorously examined and cleansed of any outliers or missing values. These refined tables were then combined, forming the foundation for the subsequent explanatory and predictive models.

Analysis:

Modeling – Explaining:

The parameter estimates indicate significance for 19 predictors, with key interest in beds, square feet, location, and zip code (Figure 1). While baths alone did not show significance, their interaction with square feet did, according to the interaction plot (Figure 2), which highlights the increased property prices with more bathrooms and larger square footage, though less so for smaller properties. The interaction between beds and square footage, however, was not significant, and adding bedrooms seemed to decrease the price, a finding that suggests additional bedrooms may not be as valued or could reduce other living spaces. These unexpected results point to potential issues with the model specification or unaccounted market factors and merit further analysis.

Checking Assumptions:

Residual Analysis:

The Ordinary Least Squares (OLS) method is used to ascertain the relationship between dependent and independent variables by minimizing the squared discrepancies between them, utilizing JMP software's diagnostic functions. The effectiveness of OLS depends on the normal distribution of residuals. Three diagnostic plots check this: one for homoscedasticity, where Figure 3 shows increasing variability, hinting at non-uniform error variances. Figure 4's Studentized Residuals Plot reveals outliers, suggesting non-normality. The Residual Normal Quantile plot in Figure 5, with its upward curve, further confirms the residuals' deviation from normality.

VIF:

The Variance Inflation Factor (VIF) was utilized to assess the presence of multicollinearity within the regression model. As the VIF value rises, the dependability of the regression outcomes diminishes. Typically, a VIF value exceeding 10 suggests substantial multicollinearity among the independent variables, warranting attention. However, according to Figure 6, all VIF values are below 7, indicating that multicollinearity is not a concern for this particular model.

Predictive Regression Model:

The study's final phase involved developing a predictive model for the pricing of the homes. The focus was on forecasting prices rather than interpreting the influence of predictors. Variance stabilizing transformations were utilized, with an optimal lambda (λ) value of -0.016 identified. The model demonstrated optimal performance near zero, as shown in Figure 7, prompting the use of a log transformation for price, creating a non-linear model. The model was trained on 75% of the dataset and validated on the remaining 25%. Its predictive accuracy was evaluated using cross-validation, as seen in Figure 8. Diagnostic checks in JMP confirmed the normal distribution of residuals, with the Residual Predicted plot (Figure 9) supporting homoscedasticity, and both the Studentized Residuals (Figure 10) and Residual Normal Quantile plots (Figure 11) indicating normality. square footage, location, and the interaction between the number of bathrooms and square footage are the most statistically significant predictors of price in this model. The number of bedrooms and ZIP codes do not have a statistically significant association with the price at the 0.05 level, although bathrooms are marginally significant and could be considered in further analysis (Figure 12).

Conclusion:

In the study, linear regression models were adeptly used for both explanatory and predictive purposes in the Dallas housing market. The explanatory model's strength lies in its use of 19 significant predictors, yielding a reliable interpretation of factors affecting house prices, as evidenced by an R-squared of 0.92. The predictive model showcased its robustness with an even higher R-squared of 0.94, as detailed in Figure 8, indicating its effectiveness in explaining 94% of the price variance. However, these models, reliant on Ordinary Least Squares (OLS), face limitations in grasping the complexities and dynamics of the real estate market. Issues like potential multicollinearity and the influence of outliers, although mitigated, might still impact the models' precision and future relevance. Therefore, while the models provide significant insights into the current market, their predictive capacity for future market trends should be approached with caution.

Figure 1: Parameter Estimates of Predictors

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	755439.12	24967.16	30.26	<.0001*
BEDS	-22266.43	7258.742	-3.07	0.0022*
BATHS	-6618.252	10166.71	-0.65	0.5152
SQUARE FEET	115.09326	11.96157	9.62	<.0001*
Group Location[<100k]	-714987.6	60597.41	-11.80	<.0001*
Group Location[>=3M]	1847376.9	58058.32	31.82	<.0001*
Group Location[1.0-1.5M]	189723.91	20591.81	9.21	<.0001*
Group Location[1.5-2M]	630916.64	39113.09	16.13	<.0001*
Group Location[2-2.99M]	1323982.7	40423.84	32.75	<.0001*
Group Location[100-200k]	-657794.4	21421.56	-30.71	<.0001*
Group Location[201-300k]	-580767.1	18511.17	-31.37	<.0001*
Group Location[301k-400k]	-527734.4	16635.08	-31.72	<.0001*
Group Location[401-500k]	-448097.8	18063.51	-24.81	<.0001*
Group Location[501-600k]	-367614.2	17056.42	-21.55	<.0001*
Group Location[601-700k]	-292204	23054.72	-12.67	<.0001*
Group Location[701-800k]	-221703.4	23481.18	-9.44	<.0001*
Group Location[801-900k]	-121991.1	28449.73	-4.29	<.0001*
Group Zip[>1.5M]	36468.638	16925.85	2.15	0.0315*
Group Zip[1-1.5M]	-11368.2	14945.8	-0.76	0.4471
Group Zip[200-300k]	-27542.26	17038.83	-1.62	0.1063
Group Zip[301-400k]	-1420.91	12078.9	-0.12	0.9064
Group Zip[401-500k]	41229.372	19437.21	2.12	0.0342*
Group Zip[501-600k]	10247.918	13340.57	0.77	0.4426
Group Zip[601-700k]	-22558.09	26811.51	-0.84	0.4004
Group Zip[701-800k]	-25920.44	18032.76	-1.44	0.1509
Group Zip[801-900k]	18254.444	18002.09	1.01	0.3108
(BATHS-2.4107)*(SQUARE FEET-2052.67)	9.1142861	2.998562	3.04	0.0024*

Figure 2: Interaction Plot of BATHS vs. Square Feet for Explaining Model

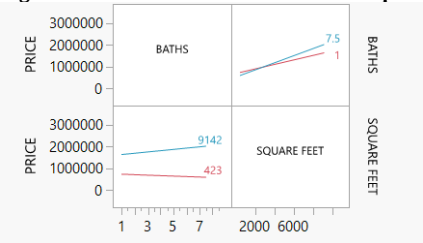

Figure 3: Residual by Predicted Plot for Explaining Model

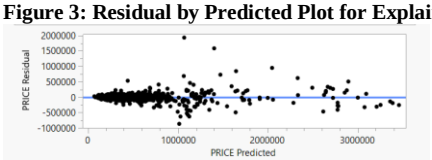

Figure 4: Studentized Residuals Plot for Explaining Model

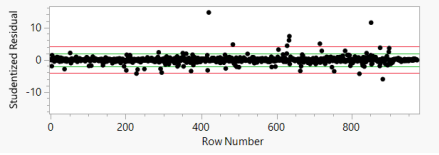

Figure 5: Residual Normal Quantile Plot For Explaining Model

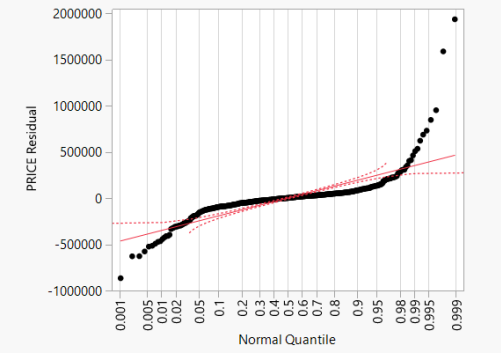

Figure 6: Parameter Estimates for Predictive Model including VIF Values

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	12.944595	0.036154	358.04	<.0001*	.
BEDS	-0.009807	0.010346	-0.95	0.3435	2.4940614
BATHS	0.0285032	0.014912	1.91	0.0564	5.1938585
SQUARE FEET	0.0001579	1.781e-5	8.87	<.0001*	8.6560285
Group Location[<100k]	-1.570261	0.093827	-16.74	<.0001*	4.638261
Group Location[>=3M]	1.2798737	0.074027	17.29	<.0001*	3.3595949
Group Location[1.0-1.5M]	0.4580058	0.029196	15.69	<.0001*	1.530431
Group Location[1.5-2M]	0.7938953	0.054442	14.58	<.0001*	2.2382262
Group Location[2-2.99M]	1.07767	0.062763	17.17	<.0001*	2.974768
Group Location[100-200k]	-1.10636	0.031961	-34.62	<.0001*	2.4502685
Group Location[201-300k]	-0.711688	0.026414	-26.94	<.0001*	2.5167411
Group Location[301k-400k]	-0.471463	0.023791	-19.82	<.0001*	1.9455953
Group Location[401-500k]	-0.263355	0.025806	-10.21	<.0001*	1.6434041
Group Location[501-600k]	-0.112406	0.024727	-4.55	<.0001*	1.6464372
Group Location[601-700k]	0.0145583	0.032165	0.45	0.6510	1.5396481
Group Location[701-800k]	0.1014487	0.035029	2.90	0.0039*	1.9216437
Group Location[801-900k]	0.2152784	0.042434	5.07	<.0001*	1.7617244
Group Zip[>1.5M]	0.0190887	0.024385	0.78	0.4340	1.4943265
Group Zip[1-1.5M]	-0.018692	0.021386	-0.87	0.3824	1.4640581
Group Zip[200-300k]	-0.067664	0.024714	-2.74	0.0063*	1.7497437
Group Zip[301-400k]	-0.015914	0.017297	-0.92	0.3579	1.438016
Group Zip[401-500k]	0.0239413	0.02828	0.85	0.3975	1.6081465
Group Zip[501-600k]	0.0157473	0.019162	0.82	0.4115	1.3657514
Group Zip[601-700k]	-0.020173	0.037011	-0.55	0.5859	2.1054701
Group Zip[701-800k]	-0.018131	0.026683	-0.68	0.4971	1.8075005
Group Zip[801-900k]	0.0646321	0.027204	2.38	0.0178*	1.6274434
(BATHS-2.4107)*(SQUARE FEET-2052.67)	-1.687e-5	4.235e-6	-3.98	<.0001*	3.1856802

Figure 7: Box Plot

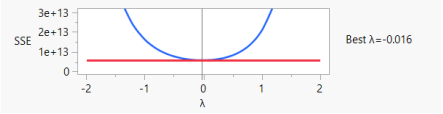

Figure 8: Cross-validation for Predictive Model

Source	RSquare	RASE	Freq
Training Set	0.9298	0.18539	705
Validation Set	0.9401	0.17080	230

Figure 9: Residual by Predicted Plot for Predictive Model

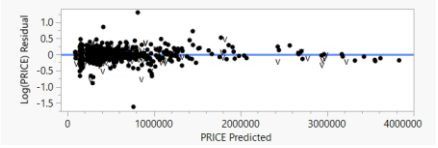

Figure 10: Studentized Residuals Plot for Predictive Model

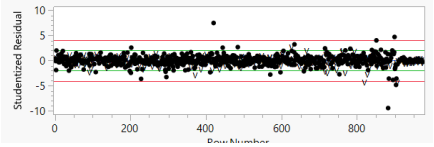

Figure 11: Residual Normal Quantile Plot For Predictive Model

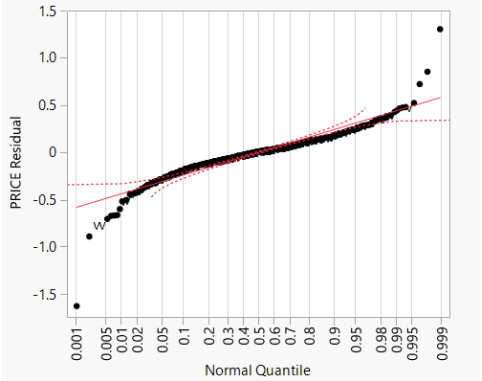

Figure 12: Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
BEDS	1	1	0.032109	0.8985	0.3435
BATHS	1	1	0.130563	3.6535	0.0564
SQUARE FEET	1	1	2.809192	78.6080	<.0001*
Group Location	13	13	60.426293	130.0673	<.0001*
Group Zip	9	9	0.552915	1.7191	0.0811
BATHS*SQUARE FEET	1	1	0.567390	15.8769	<.0001*