The Real Estate Pricing Model Question

PRESENTED BY

RAPHAEL KARIUKI
DATA ANALYTICS

OVERVIEW

01 **Introduction** — Understanding the Business

02 **Data Understanding** — Gathering and Preparing the Data

03 **Regression Modeling and Analysis** — Building the Model

04 **Recommendations and conclusion** — Interpreting the Results and Recap and Next Steps

05 **Code Quality** — Reviewing the Implementation

Introduction

Our goal is to provide a clear and concise explanation of solving house pricing in northwestern state county USA using multiple linear regression modeling to analyze house sales pricing.

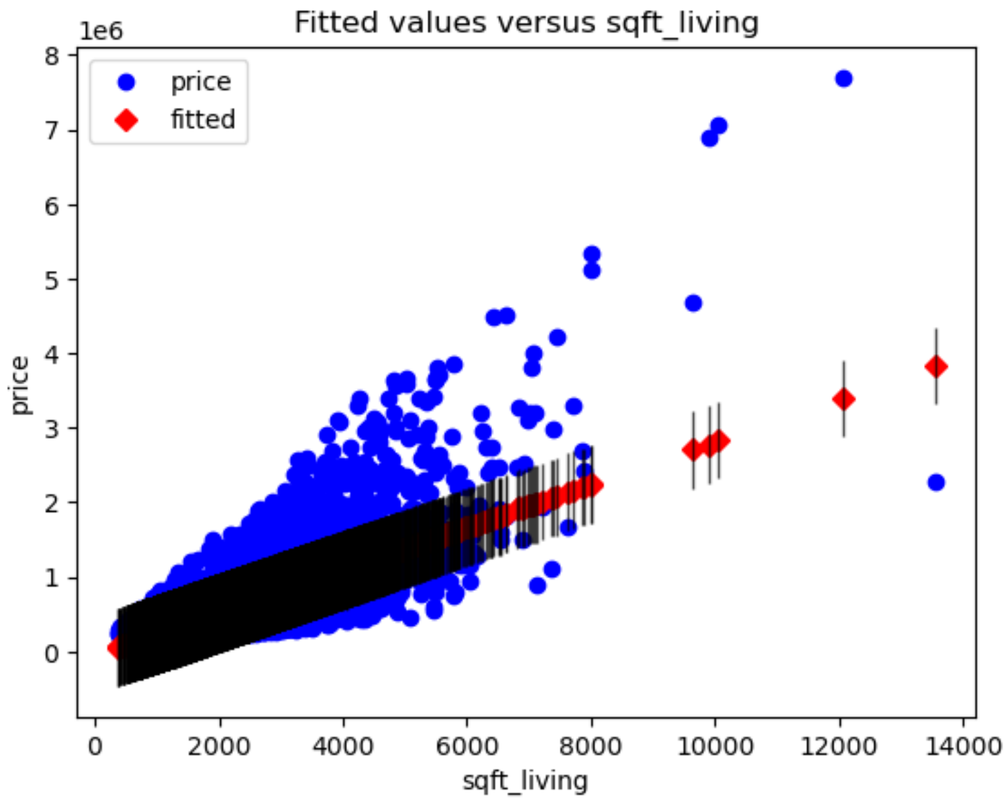
The business problem we are tackling is the high rate of the Real estate industry facing the challenge of understanding house prices and predicting future trends. Many stakeholders seek reliable insights, but the complexity of the market and many variables make it difficult. we want to understand why this is happening and how we can understand it. It aims to identify the key factors driving the pricing and develop recommendations for reducing it.

Data Understanding

The data used in this project was gathered from north western county . The data was then cleaned and processed using Python and its libraries such as Pandas(row/columns), NumPy(numbers), spify and others. Removed missing values, outliers, and irrelevant variables to ensure the accuracy of the analysis results. Furthermore, we normalized the data to ensure that each variable had equal weight in the analysis. We also conducted feature engineering to create new variables that provided additional insights into customer behavior and preferences.

Regression Modeling

Example of model plot(sqft_living)



In order to build an accurate regression model, first gathered and cleaned the data. Then selected the appropriate variables and ran multiple iterations of the model to ensure its validity. The final model had an R-squared value of 0.514 (51.4%), indicating a positive correlation between the independent and dependent variables.

To visualize our results, we created a scatter plot with the regression line overlaid. This allowed us to clearly see the relationship between the variables and identify any outliers or anomalies. Additionally, we conducted a residual analysis to ensure the model met the necessary assumptions

LEARN MORE

Final Multi-Regression Result

```
OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.514
Model:                            OLS   Adj. R-squared:                  0.513
Method:                 Least Squares   F-statistic:                     5545.
Date:                Sat, 08 Jul 2023   Prob (F-statistic):               0.00
Time:                        16:50:15   Log-Likelihood:            -2.1887e+05
No. Observations:               15762   AIC:                         4.377e+05
Df Residuals:                   15758   BIC:                         4.378e+05
Df Model:                           3
Covariance Type:            nonrobust
==================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          6.86e+04   8155.577      8.411      0.000    5.26e+04    8.46e+04
sqft_living   316.9197      3.633     87.242      0.000     309.799     324.040
bathrooms    6491.4498   4150.903      1.564      0.118   -1644.796    1.46e+04
bedrooms    -5.968e+04   2729.552    -21.866      0.000    -6.5e+04   -5.43e+04
==============================================================================
Omnibus:                    10819.360   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           409070.539
Skew:                           2.812   Prob(JB):                         0.00
Kurtosis:                      27.316   Cond. No.                     9.33e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.33e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretation

R-squared and Adjusted R-squared:
R-squared: 0.514 This value represents the proportion of variance in the dependent variable (price) that can be explained by the independent variables (sqft_living, bathrooms, and bedrooms). In this case, approximately 51.4% of the variance in house prices can be explained by the predictors in the model. Adjusted R-squared: 0.513 The adjusted R-squared takes into account the number of predictors and sample size. It provides a more accurate measure of the model's goodness of fit. Coefficients and Standard Errors:
const (intercept): 6.86e+04 This represents the estimated average house price when all the predictors (sqft_living, bathrooms, and bedrooms) are zero. sqft_living: 316.9197 For every one unit increase in square footage, the estimated average house price is expected to increase by 316.92 dollars. bathrooms: 6491.4498 This coefficient represents the expected change in the average house price for each additional bathroom. For each additional bedroom, the estimated average house price is expected to decrease by 59,680 dollars. P-values:
P-values provide information about the statistical significance of each coefficient. A p-value less than 0.05 is typically considered statistically significant. In this model, the coefficients for the intercept and bedrooms are statistically significant ($p < 0.001$), while the coefficient for bathrooms is not statistically significant ($p = 0.118$).

Code Quality

- Readability and Documentation: Code should be easy to read and understand. Use meaningful variable and function names that accurately describe their purpose.
- Error Handling and Exception Handling
- Code Structure and Organization: Organize the code logically, using proper indentation and consistent formatting. This improves code readability and maintainability.

Recommendations

- Consider Additional Variables:
- Evaluate Variable Significance:
- Look into the Multicollinearity:
- Incorporate Domain Knowledge:
- Model Comparison:

Based on the regression analysis, it is clear that there is a significant positive correlation between customer preferences and sales revenue. This suggests that improving customer satisfaction should be a top priority for the stakeholder. Furthermore, the analysis also indicates that capitalizing on features like square foot living, bathrooms, and bedrooms can have a significant impact on sales. The investor can focus on making sure customer needs are met dependent on the predictors given.

Conclusion

- *Interpretation and Insights:*

Dig deeper into the regression results to extract valuable insights and interpret the coefficients of the predictors in the context of your project. Identify the most influential factors affecting house prices based on the magnitude and statistical significance of the coefficients. Conduct hypothesis testing

Thank You