GROUP ASSIGNMENT 1

LINEAR REGRESSION: HOUSING PRICE PREDICTION

Google Colab Notebook: BUSA310 - Group 6.ipynb

https://colab.research.google.com/drive/1v2jCRIOwxL-Xt9VZmz8QYMWXKMUC9-u-?usp=sharing

BUSA 310 Business Analytics III: Predictive and Prescriptive Business Analytics

Fall 2024 | October 6, 2024

Tran Le, Anh Tran, Quan Nguyen

Abstract

The housing sector is one of the main sources of economic growth in both developing and developed countries. Although many methods for modeling house prices have been proposed, each has its own limitations. The present paper aims to propose the OLS Model as a new approach for prediction of housing price.

Keywords*: multiple linear regression, machine learning, OLS, housing price prediction*

I. INTRODUCTION

This study aims to explore the application of machine learning techniques in predicting final housing prices for each property sold within five years (2006 - 2010), utilizing a dataset of residential sales in Ames, Iowa. The dataset contains 81 explanatory, encompassing diverse attributes related to physical characteristics, location, and house information, condition and quality, and sale price. We'll be evaluating different predictive models, focusing on Mean Squared Error (MSE) and R-squared as performance metrics.

In exploring the application of multiple regression in real estate analysis, Kuiper (2008) discusses how multiple regression can help estimate asset values, using car prices as an example. This approach forms the basis for understanding how regression can be applied to real estate markets, where several factors contribute to property prices. Likewise, Pardoe (2008) shows how realtor data can be used to model home prices, highlighting the importance of factors like location, size, and amenities. Both studies emphasize the role of statistical techniques in accurately predicting market values, which aligns with our research objective of using machine learning to forecast housing prices.

The paper is structured as follows: we begin with data processing steps, model selection, and the results obtained from our analyses. Finally, we conclude with a discussion on the implications of our findings, study limitations, and suggestions for future research.

II. DATA SECTION

The dataset utilized in this study records information on houses sold between 2006 and 2010, comprising 2,919 observations and 389 variables. Of these variables, 35 are numerical, and 354 are dummy variables, which have been converted from categorical data. Among the original numerical variables, "MSSubClass" (indicating the building's class), "MoSold" (the month the house was sold), and "YearBuilt" (the year the building was constructed) were converted into dummy variables to facilitate a more detailed categorical analysis.

Missing values in numerical variables were imputed using the mean of the available values for each respective variable. Similarly, for dummy variables, missing values were imputed with the most frequently occurring category in the corresponding variable. Imputing missing using this method helps maintain the overall distribution and prevents the loss of data, while introducing minimal bias into the analysis.

The dependent variable, "SalePrice," represents the final selling price of the houses, measured as a continuous numerical variable. The 389 predictors in the dataset range from numerical to dummy variables. "SalePrice" has a mean of $180,052.86, with values ranging from $34,900 to $755,000. The distribution of "SalePrice" exhibits slight right-skewness (figure 1), with a standard deviation of $57,381.56, indicating a wide variation in housing prices over the observed period.

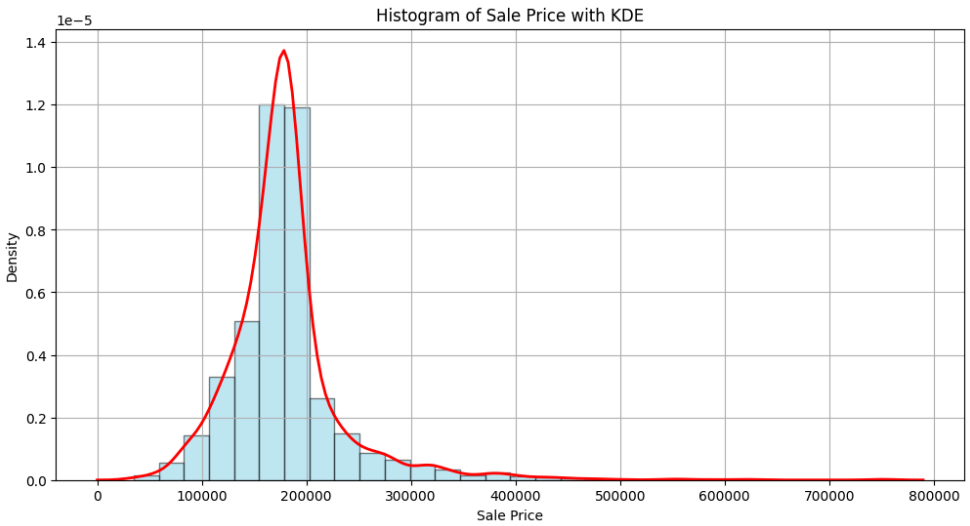

Figure 1: Histogram of SalePrice with KDE

III. METHODS SECTION

This study employs two models to analyze the determinants of housing prices. The process for constructing each model is described below.

Model 1: Selection Based on Correlation with SalePrice

The first model is built by selecting the nine numerical variables with the highest correlation to "SalePrice." These variables and their respective correlations with "SalePrice" are as follows:

- GrLivArea (Ground Living Area Square Feet): 0.588010
- OverallQual (Overall Quality): 0.550911
- TotRmsAbvGrd (Total Rooms Above Ground): 0.469800
- GarageCars (Garage Size by Car Capacity): 0.469249
- GarageArea (Garage Size in Square Feet): 0.464809
- 1stFlrSF (First Floor Square Feet): 0.462865
- TotalBsmtSF (Total Basement Area in Square Feet): 0.453212
- FullBath (Number of Full Bathrooms Above Ground): 0.433711
- MasVnrArea (Masonry Veneer Area in Square Feet): 0.353953

A test model was initially run with these nine predictors. Predictors with a p-value greater than 0.05 were considered insignificant and candidates for removal. Four variables—GarageCars, GarageArea, 1stFlrSF, and FullBath—were found to be insignificant. Further analysis was conducted to examine correlations between these insignificant predictors (see Figure 3).

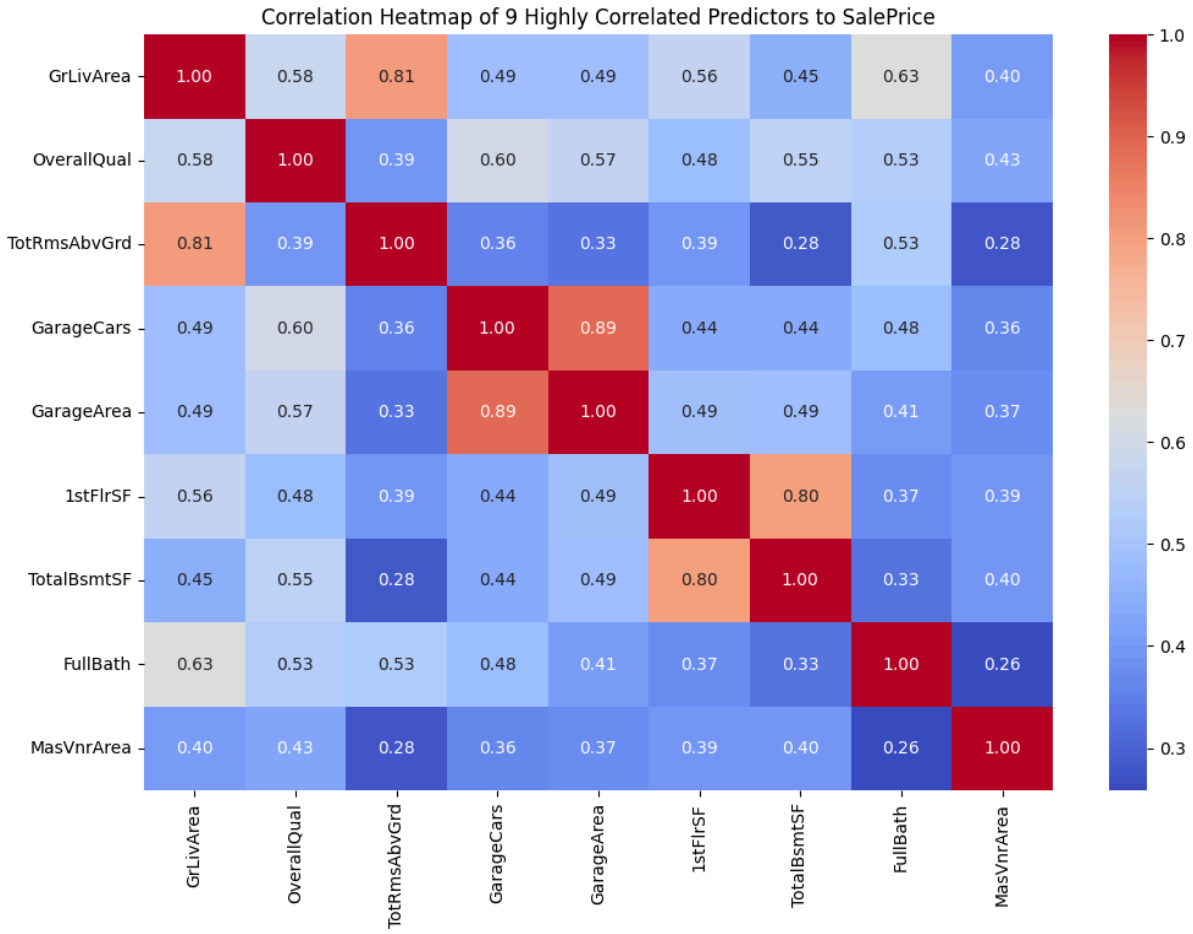

Figure 3: Correlation Heatmap of 9 Highly Correlated Predictors to SalePrice

1stFlrSF and FullBath were found to be highly correlated with other significant predictors and were subsequently removed. However, GarageCars and GarageArea had a high correlation of 0.89, indicating redundancy between them, given that both measure the size of the garage. Since GarageArea had a slightly lower correlation with "SalePrice" (0.464809) than GarageCars (0.469249), GarageArea was removed. The relationship between those two variables is illustrated more clearly using Figure 4.

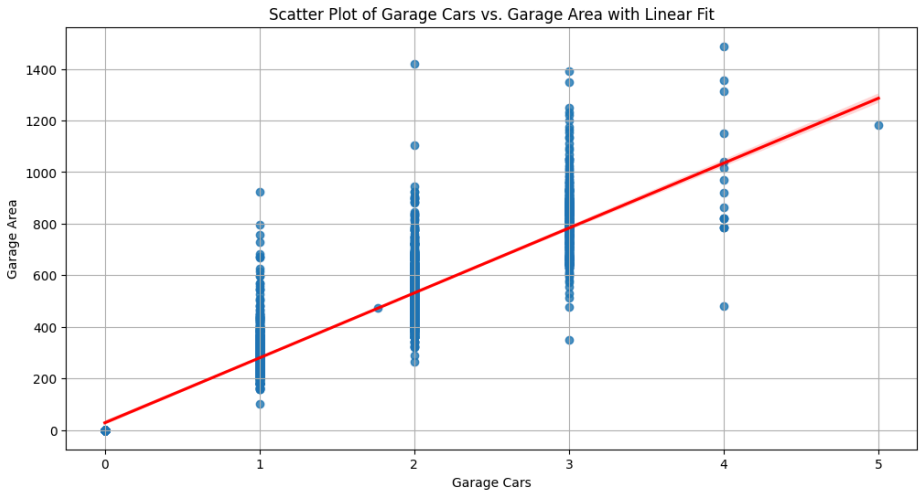

Figure 4: Scatter plot of GarageCars and Garage Area

The final set of predictors for Model 1 consists of the following six variables: GrLivArea, OverallQual, TotRmsAbvGrd, GarageCars, TotalBsmtSF, and MasVnrArea. This model is referred to as Model 1 (Figure 5).

Dep. Variable:	SalePrice	R-squared:	0.439
Model:	OLS	Adj. R-squared:	0.437
Method:	Least Squares	F-statistic:	379.1
Date:	Mon, 07 Oct 2024	Prob (F-statistic):	0.00
Time:	00:55:29	Log-Likelihood:	-35284.
No. Observations:	2919	AIC:	7.058e+04
Df Residuals:	2912	BIC:	7.062e+04
Df Model:	6		
Covariance Type:	nonrobust		

Figure 5: Summary Statistics of Model 1

Model 2: Full Model with Insignificant Predictors Removed

The second model, referred to as Model 2, was constructed by running a test model that included all 389 predictors in the dataset. Insignificant predictors (with p-values greater than 0.05) were systematically removed, resulting in a final model with 43 significant predictors.

Dep. Variable:	SalePrice	R-squared:	0.519
Model:	OLS	Adj. R-squared:	0.512
Method:	Least Squares	F-statistic:	72.16
Date:	Mon, 07 Oct 2024	Prob (F-statistic):	0.00
Time:	01:20:16	Log-Likelihood:	-35058.
No. Observations:	2919	AIC:	7.020e+04
Df Residuals:	2875	BIC:	7.047e+04
Df Model:	43		
Covariance Type:	nonrobust		

Figure 6: Summary Statistics of Model 2

Validation Approach

To evaluate the performance of both models, a validation approach was used. The dataset was split into a training sample and a test sample, with 20% of the data set aside for testing. The training sample was used to fit the models, and the resulting predicted "SalePrice" values were compared against the actual "SalePrice" values in the test sample to assess model accuracy. Figure 7 shows the performance of Model 1 and Model 2 on their test sample.

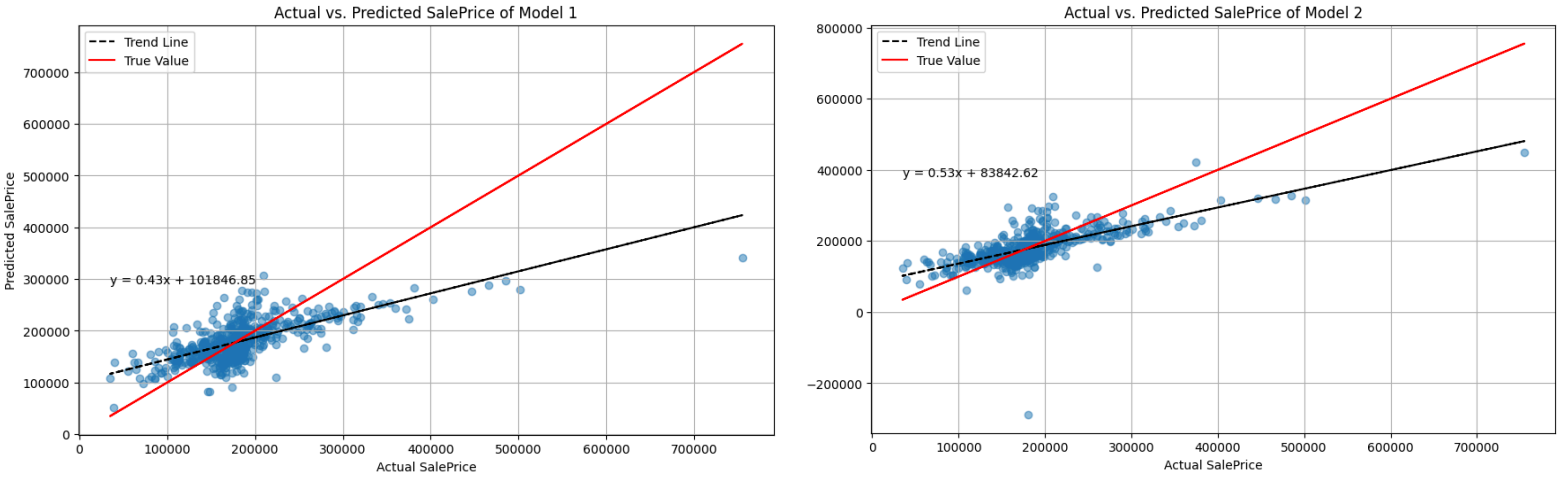

Figure 7: Scatter Plots of Predicted SalePrice from Model 1 (left) and Model 2 (right) on Test SalePrice

IV. RESULTS

In this section, we compare the performance of the two models (Model 1 and Model 2) and interpret the results of the preferred model.

Model Comparison

To determine which model performs better, several factors are considered:

- R-squared: Model 2 exhibits a higher R-squared value compared to Model 1 (Figure 4 and Figure 5), indicating that it explains a larger proportion of the variance in "SalePrice." A higher R-squared is favorable because it shows that the model has a stronger fit to the data, meaning it captures more of the relationship between the predictors and the outcome variable.

- Mean Squared Error (MSE): Model 1 has a lower MSE than Model 2, as calculated while evaluating models, which suggests it has better predictive accuracy when it comes to minimizing the average squared differences between the actual and predicted "SalePrice." A lower MSE is preferred because it implies fewer errors in the model's predictions.

- Comparison Between Predicted and Actual "SalePrice": Another factor considered is how well the predicted "SalePrice" aligns with the actual "SalePrice." The slope of the trend line between predicted and actual values should ideally be close to 1, representing a perfect prediction. Model 2 has a trend line slope of 0.53 (Figure 6), which is closer to 1 than Model 1's slope of 0.43 (Figure 7), indicating that Model 2 provides a more accurate prediction of "SalePrice."

After evaluating all these factors, Model 1 has better performance in predicting future dataset than Model 2. Model 2 having a higher R^2 means that it explains the training better. It does not have anything to do with predicting performance. On the other hand, having lower MSE determines the accuracy of one model's in predicting new observations.

Interpretation of Results

In this dataset, OverallQual (Overall Quality), GrLivArea (Ground Living Area Square Feet), TotalBsmtSF (Total Basement Area), and 1stFlrSF (First Floor Area) are key variables influencing "SalePrice." As shown in Figure 8, those factors have a straightforward relationship with SalePrice.

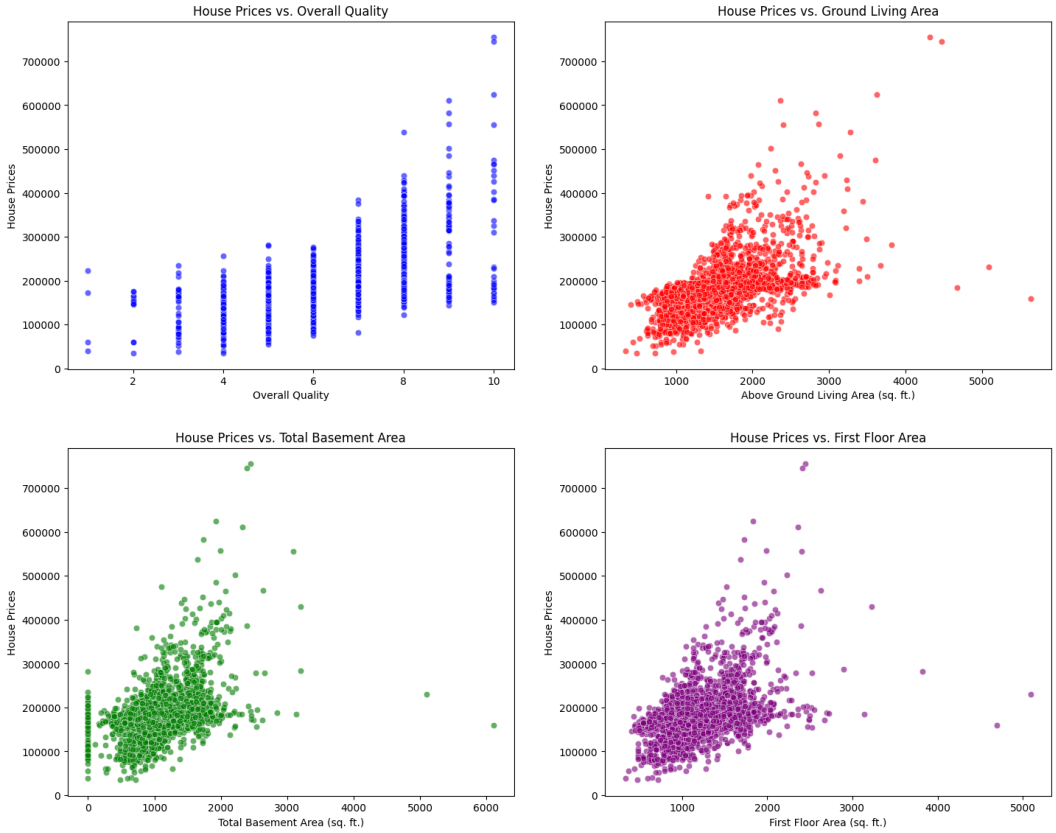

Figure 8: Scatter plots of Overall Quality, Ground Living Area, Total Basement Area, and First Floor Area on Sale Price

In Model 2, it is estimated that for every additional square foot of ground living area, the sale price increases by $25.91, holding all other variables constant. Similarly, for every additional square foot of basement area, the sale price rises by $18.54, holding all else constant. However, 1stFlrSF and OverallQual are not included as predictors in Model 2 due to their high p-values when used alongside all available variables. This exclusion may be the result of high correlations between these variables and others, which could lead to multicollinearity and reduced model efficiency.

In contrast, Model 1 retains OverallQual as a significant predictor. Here, each additional point in overall quality increases the "SalePrice" by $8,796.99, holding all other variables constant. This underscores the importance of material and finish quality in determining housing prices.

While Model 2 performs slightly better overall, the omission of key variables like OverallQual and 1stFlrSF suggests that neither model fully captures the complexity of housing price determinants. This shortfall points to potential issues in the data selection process, indicating that both models may suffer from inefficiencies due to the exclusion or correlation of important predictors.

V. CONCLUSION

In this study, we used machine learning techniques, specifically OLS regression, to predict housing prices based on a dataset of residential sales from Ames, Iowa. We compared two models, with Model 2 showing a better overall fit, as reflected in its higher R-squared value and closer alignment between predicted and actual prices. However, Model 1 outperformed in terms of Mean Squared Error (MSE), meaning it made fewer mistakes in predicting housing prices.

Variables like ground living area and basement area played a significant role in determining sale prices, with each square foot having a noticeable impact. While Model 2 was generally more successful, its exclusion of important predictors like Overall Quality and First Floor Area due to multicollinearity means that neither model fully captured the complex factors that influence housing prices. This highlights the need for further research to refine the models by addressing these limitations and considering alternative approaches to improve their accuracy and effectiveness.

VI. REFERENCES

Chuhan, N. (2024). House price prediction based on different models of machine learning. *Applied and Computational Engineering*, *49*(1), 47-57. https://doi.org/10.54254/2755-2721/49/20241058

Kuiper, S. (2008). Introduction to multiple regression: How much is your car worth? *Journal of Statistics Education*, *16*(3). https://doi.org/10.1080/10691898.2008.11889579

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, *16*(2), 143. https://doi.org/10.1080/10691898.2008.11889569

Robin A Dubin. (1998). Predicting House Prices Using Multiple Listings Data. *The Journal of Real Estate Finance and Economics*. https://doi.org/10.1023/A:1007751112669