Housing Price Analysis

Stella Ramirez, Liuqian Bao, Andrew Cheng

13 December, 2023

Introduction

The following is a summary of the analysis we conducted on the Ames, Iowa, residential housing data. Through model building and accuracy testing, we have provided several methods through which to compute and predict the sale price of a property, which a client can then apply to their needs.

The citation for our original data source is: Anna Montoya, DataCanary (2016). House Prices - Advanced Regression Techniques. Kaggle. Obtained from <https://kaggle.com/competitions/house-prices-advanced-regression-techniques>

The main questions of interest we had were which variables significantly influence a properties price, and how these may help us predict the price of a property. Therefore, our response, or dependent variable was “SalePrice”, the property's sale price in dollars. Our dataset provided us with many predictor variables, both categorical and quantitative. However, as we worked, we used various methods to identify the most influential variables, and to minimize the noise created by irrelevant information.

Exploring a Single Predictor

After investigating each provided variable individually(cf. Step 1), we noticed a strong association between the finished basement square footage(variable name: BsmtFinSF1) and the sale price. In order to explore this hypothesis, we constructed a linear regression model in which BsmtFinSF1 was the only predictor variable. The following shows the equation we created:

$$log(y) = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$log(SalePrice) = 11.84 + 3.627 \times 10^{-4} BsmtFinSF1$$

This model shows that with an increase in finished basement square footage, the log Sale Price will increase as well. It is important to note that after analyzing the naive linear model in which the response was simply SalePrice, we found that several assumptions needed for an accurate linear model weren't fulfilled. Namely, the variance was not constant, and the points did not closely follow a normal distribution. While normality is not essential for predictions, given that we are working with a large amount of data, we transformed the SalePrice with a log function to address the constant variance. Upon doing so, the residual models reveal that the points now have more constant variance, and a qq-plot displays that this model fulfills normality(cf. Step 2, Page 2-3).

Now that we had created the model, we performed a hypothesis test to see if BsmtFinSF1 is in fact a significant predictor. Because the p-value returned, we were able to reject the null Hypothesis($\beta_1 = 0$), and accept the alternative hypothesis that $\beta_1 > 0$, confirming that BsmtFinSF1 and SalePrice are positively correlated. Because the goal is to view how our predictors change the sale price of a property, we constructed a confidence interval for the average SalePrice(logged), with finished basement area equal to the average of BsmtFinSF1 in the data. We concluded that with 95% confidence, this value is estimated to be between 11.97263 and 12.03798. While this may be helpful, we looked at the r^2 value and found that this model only explains 14.83% of the variation in the log of SalePrice. Now that we'd explored how BsmtFinSF1 may affect the response when it is the sole predictor, we decided to construct a model that conveyed more information.

Constructing a More Accurate Model

In order to more accurately predict the SalePrice of a property, we performed both backward and forward selection, keeping in mind that the response is the log of the SalePrice. These methods allow us to select only the most significant predictors. Both of these methods determined that these are external quality(ExterQual), LotArea, GarageCars, finished basement square footage (BsmtFinSF1), HalfBath, FullBath, kitchen quality (KitchenQual), and open porch square footage (OpenPorchSF). The following is the resulting linear equation:

$$log(y) = \beta_0 + \sum_{i=1}^{8} \beta_i x_i + \varepsilon$$

$$log(SalePrice) = 11.78 + 4.491 \times 10^{-6} LotArea + 0.108 GarageCars + 1.950 \times 10^{-4}$$
$$BsmtFinSF1 + 0.100 HalfBath + 0.1725 FullBath + 4.452 \times 10^{-4} OpenPorchSF +$$
$$- 0.8778 I_{\{ExterQual=Fa\}} - 0.1905 I_{\{ExterQual=Gd\}} - 0.3509 I_{\{ExterQual=TA\}} - 0.2845 I_{\{KitchenQual=Fa\}}$$
$$- 0.0799 I_{\{KitchenQual=Gd\}} - 0.1889 I_{\{KitchenQual=TA\}}$$

This equation tells us that an increase in lot area, garage cars, finished basement square footage, half bath, full bath, or open porch will result in an increase in the log of SalePrice, and SalePrice will decrease if the categorical variables external quality and kitchen are not excellent. With this equation, the r^2 value revealed that 81.35% of the variation in the response was explained. Therefore, it is a much more useful model. As a client, you may consider using this equation to estimate or predict future SalePrice values.

Ensuring We Have Used Variables of Influence

We performed two regression techniques, LASSO and Ridge, attempting to ensure that none of these eight variables were irrelevant, or too closely correlated with another that they incorrectly estimated the response. After using these methods to find the best models, we constructed this graph to compare the three.

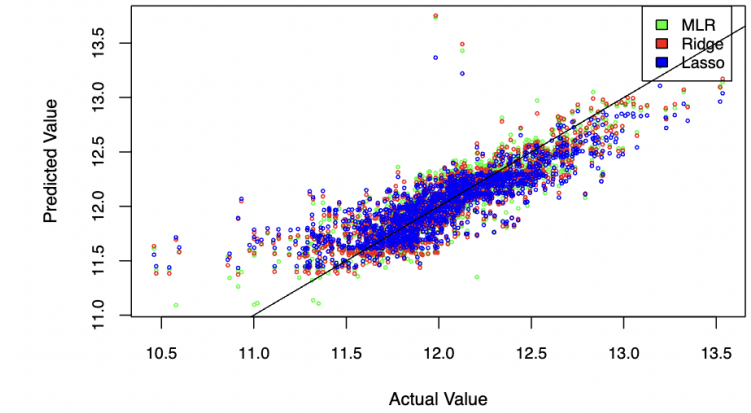

Figure 6: scatter plot of the predicted vs. actual values for MLR, Rridge, and Lasso Models

As you can see, the three equations result in very similar results, indicating that our original MLR was likely a good fit. In summary, the MLR, Ridge, and LASSO models have R-squared values of 0.828, 0.7132, and 0.6972, respectively, suggesting that MLR explains the most of the response variance. However, by analyzing the predicted vs. response graph, Ridge and Lasso have lower variances in predictions and might generalize better to unseen data due to their reduced complexity.

Innovation

When we explored residual plots, we found that our data had heteroscedasticity, or unequal

variance. In order to fix that, we took the log of our response variable. Another way in which this issue can be addressed is through weighted linear regression, or the method of weighted least squares. This method places weights on observations so that the ones with smaller error variance are more influential.

The model equation we obtained after adding weights is:

$$SalePrice = 2.255 \times 10^5 + 1.911 LotArea + 1.318 \times 10^4 GarageCars + 31.43$$
$$BsmtFinSF1 + 1.488 \times 10^4 HalfBath + 2.363 \times 10^4 FullBath + 31.30 OpenPorchSF +$$
$$- 1.190 \times 10^5 I_{\{ExterQual=Fa\}} - 7.269 \times 10^4 I_{\{ExterQual=Gd\}} - 9.836 \times 10^4 I_{\{ExterQual=TA\}}$$
$$- 6.288 \times 10^4 I_{\{KitchenQual=Fa\}} - 4.149 \times 10^4 I_{\{KitchenQual=Gd\}} - 5.911 \times 10^4 I_{\{KitchenQual=TA\}}$$

Comparing the summaries of our unweighted and weighted models, we observed changes in coefficient estimates, standard errors and, r^2 values. Especially, the residual standard error shows a drastic change from the model without weights. This means that the values that are predicted with the weighted model are much more accurate and aligned with the actual observations. Also, because the r^2 value increased, we know that the weighted model is able to explain more of the variance in SalePrice.

Limitations

For the purposes of our linear regression models, we omitted variables that are time dependent in the variable selection process. However, just by common sense, the sale price of a property could have some correlation with the year in which it is built or renovated. We can obtain more information about the housing market by further investigating time-dependency of the sale price using some time series models in the future.

Our analysis assumes a linear relationship between the response and the predictors, which might not be entirely true because the real housing market can be very complex. It is also very difficult to detect non-linear relationships with a large number of categorical variables. Thus we might need additional tools to take into account non-linear relationships, and gain a more holistic understanding of the housing price. Also, the dataset contains outliers that could potentially influence the model fit and make the predictions inaccurate.

Conclusion

In conclusion, we found that the most statistically significant variables from our dataset are external quality(ExterQual), LotArea, GarageCars,finished basement square footage (BsmtFinSF1), HalfBath, FullBath, kitchen quality (KitchenQual), and open porch square footage (OpenPorchSF). As discussed, increases/decreases in each of these variables will impact the SalePrice. Due to this conclusion, we suggest clients use these as factors to be considered when deciding which properties to invest in, or how to price a property for the market. In addition, the most accurate model is the one created using the weighted least squares method, and therefore is the one that should be fitted when making predictions.