Aaron Niecestro
April 30, 2024
PH1965 Bayesian Data Analysis

Washington D.C. Housing Model using Bayesian Logistic Regression

Abstract

Predicting whether a house or property is worth the price specified on property websites is very difficult. Many frequentist model approaches have explored this type of analysis but not so many using the Bayesian model approach. After extensive cleaning of the data, the data was broken into a training set (80%) and a testing set (20%), and variable regularization and variable selection were performed. Ridge alone, Horseshoe alone, and Horseshoe *and* Ridge in combination were used to perform variable regularization while Lasso regression was used to perform variable selection. The model did have high accuracy, but more analysis needs to be conducted to determine how to increase the overall model accuracy.

Introduction

Predicting whether a house or property is worth the price specified on property websites is very difficult. Many frequentist model approaches have explored this type of analysis, but taking the variance into account using a Bayesian approach is not done as often or much since the required computing power is very high. I have experience in this since previous analyses have been completed in the past using frequentist and machine learning models; however, the predictive accuracy for some of these statistical methods like frequentist logistic regression, KNN, principal competent analysis, and classification trees was less than seventy percent. This study aims to see whether a Bayesian approach might result in higher predictive accuracy, especially one greater than eighty percent. The goal of this project was to create a Bayesian logistic regression model to determine which variables in a Washington, DC housing model determine whether a property is qualified as an indicator to reflect if a sale price is representative of market value according to the office's internal criteria.

Data Description

The data is about residential property descriptions in Washington, DC USA, and addresses information about housing prices, geography, property type, and much more using the Geographic Information System from calendar year 2018. During the preliminary data description, it was discovered that there was a lot of missing data. Before the data cleaning process began, there were 158957 observations with 49 columns, but after the cleaning process finished, there were 57610 with 49 columns. The data were cleaned in a way such that all missing data rows were deleted since, if imputation was used for continuous variables, the results would be too skewed (as we see in Table 1 below, the standard deviation is very high for most variables). Usually, the standard deviation should decrease after ten thousand observations, but that was not the case for the property data in Washington, DC.

The variables that were used in the model analysis process utilized two criteria. The first criterion is that after an extensive literature review, they had to be found to be important in

housing, property, and apartment models. From our data using the first criterion, it was found that thirteen variables were needed for our model analysis. The second criterion is that the standard deviation cannot exceed half of the mean for a continuous variable and that the categories need to be evenly distributed with no more than two categories being underrepresented. From these two joint criteria, it was ultimately shown that only ten variables were needed for our model analysis.

Proposed Approach-Model Specification

To approach this problem, a Bayesian logistic regression model with ten predictor variables using a non-informative prior was utilized. These ten predictor variables were as follows: price per $100,000, bathroom, rooms, bedrooms, stories, fireplaces, kitchens, years since the last amenity improvement, years since the place was remodeled, and air conditioning. The non-informative prior used was a normal distribution with a mean of zero and a variance of six. The reason that the variance was six was because this was the average of the seven most important variables (price per $100,000, bathroom, rooms, bedrooms, stories, fireplaces, and air conditioning) you find in most housing models and websites. A training set consisting of eighty percent of the data and a testing set consisting of twenty percent of the data were created before running any model analysis. Cross-validation was used for the model analysis since this was the best-known way to get the most accurate and reliable results.

Data Analysis Methodology

After the data cleaning and summary statistics (mean, median, Interquartile ranges, standard deviation, etc.) for the training and testing sets were completed, the next step was the model analysis. Since the variables I wished to use to analyze my data had been chosen and the prior distribution decided upon, the next step was to analyze which variables to use in my final model. I used a package in R called “bayesreg” to perform ridge alone, horseshoe alone, and horseshoe *and* ridge in combination to perform regularization, and lasso regression to perform variable selection for the Bayesian logistic regression. The regularization methods gave me the same results that stated I should not use the remodeling age and number of bathroom variables. However, I decided to use and follow the results from the lasso regression method since lasso performs variable selection, and it showed me only to eliminate the remodeling age variable, which had the highest standard deviation.

Once all the variable selection for the final model was completed, I created an R program to run my model analysis and an R stan program to perform the Bayesian logistic regression model using cross-validation. I only used one chain and 10,000 iterations with cross-validation to create the final model. After running the 10,000 iterations of the model, I checked the test performance and made sure that the iterations converged using trace plots. Afterward, I checked the coefficients and their standard deviations to ensure nothing stood out. As one can see in Figures 1 and 2 below, the credible interval for the intercept and the respective coefficients are not only not that large but they are very small.

The model results are not that surprising since the key variables like price, air conditioning, and bathrooms give positive odds while the others give negative results. For every additional hundred thousand dollar increase in price that a property in DC is worth, the odds of that property being qualified as an indicator to reflect if a sale price is representative of market

value according to the office's internal criteria increase by 15.6% while considering all other variables being held constant. For every additional bathroom, property in DC has the odds of that property being qualified as an indicator to reflect if a sale price is representative of market value according to the office's internal criteria increase by 3.2% while considering all other variables being held constant. If a property has air conditioning, the odds of that property being qualified as an indicator to reflect if a sale price is representative of market value according to the office's internal criteria increase by 60.42% while considering all other variables being held constant.

As one can see in Figure 3 below, the final model accuracy was 79.185% ± 0.025%, which is not optimal since the project aimed to have an accuracy greater than 80%; this might have been expected since only cross-validation was used. While not perfect, obtaining an accuracy over 50% using a housing model is very good, especially when utilizing a Bayesian approach.

Conclusion

I conclude that although my project's final result is not optimal since the accuracy was less than 80%, it is still a useful model to use for determining whether a property is qualified as an indicator to reflect if a sale price is representative of market value according to the office's internal criteria. This is because of the following three reasons. First, only cross-validation using training and testing sets were used for the model analysis with a set seed of 400. Using this approach, the results can change depending on how you randomly choose the training and test observations. It would have been better to use Leave-One-Out cross-validation, but I did not know how to implement this, so I utilized the next best step and used the training and testing sets approach. Second, I did not use many categorical variables since I determined that the variables in my dataset were not that relevant. Maybe if I used different variables and allowed Lasso regression to perform variable selection, then the accuracy might have increased. Third, I did not use a random intercept and slope model, which might have been more appropriate since there could be variation in the intercept and variables especially when you consider the specific property location/ward in Washington, DC. Overall, and as alluded to above, the model may not be the best but it does do a sufficient job in addressing the objective of this project.

Appendix

Table 1: Demographics

Variable	Qualified (N = 45767)	Not Qualified (N = 11837)
Price (per $100,000)	6.291 (5.45)	3.867 (5.254)
Bathroom	2.592 (1.121)	2.305 (1.164)
Rooms	7.436 (2.223)	3.404 (2.559)
Bedrooms	3.425 (1.092)	2.055 (1.189)
Stories	2.103 (0.424)	1.35 (0.442)
Kitchens	1.229 (0.591)	1.35 (0.793)
Fireplace	0.686 (0.925)	0.4865 (0.827)
Improvement Years	47.89 (16.965)	54.07 (15.763)
Remodel Years	8.494 (11.62)	8.01 (12.648)
Air Conditioning		
Yes	35234 (76.99%)	5077 (42.89%)
No	10533 (33.01%)	6760 (57.11%)

Footnote: Continuous Variable: Mean (standard deviation); Categorical Variable: Frequency (Percentage).

Table 2: Bayesian Logistic Regression Models

Variable	Coefficient	95% Credible Interval	Odds	95% Credible Interval for Odds
Intercept	1.6175 (0.0975)	(1.5502, 3475.135)	5.0405 (1.1024)	(4.1917, 6.1155)
Price (per $100,000)	0.1449 (0.0048)	(0.1355, 0.1544)	1.156 (1.0048)	(1.1452, 1.167)
Bathroom	0.0317 (0.0193)	(-0.0057, 0.0691)	1.0322 (1.0195)	(0.9943, 1.0716)
Rooms	-0.0399 (0.0086)	(-0.0565, -0.0232)	0.9608 (1.0086)	(0.9451, 0.9771)
Bedrooms	-0.0995 (0.0157)	(-0.1304, -0.0687)	0.9053 (1.0158)	(0.8778, 0.9336)
Stories	-0.0804 (0.0315)	(-0.1427, -0.0202)	0.9228 (1.032)	(0.867, 0.98)
Kitchens	-0.0977 (0.0237)	(-0.1437, -0.051)	0.9069 (1.024)	(0.8661, 0.9503)
Fireplace	-0.0746 (0.0167)	(-0.1084, -0.0425)	0.9281 (1.0169)	(0.8973, 0.9584)
Improvement Years	-0.0077 (0.0009)	(-0.0095, -0.0059)	0.9924 (1.0009)	(0.9905, 0.9942)
Air Conditioning = Yes	0.4726 (0.0288)	(0.4154, 0.5295)	1.6042 (1.0292)	(1.515, 1.6981)

Footnote: Mean (standard deviation).

Figure 1: Intercept and the Credible Interval

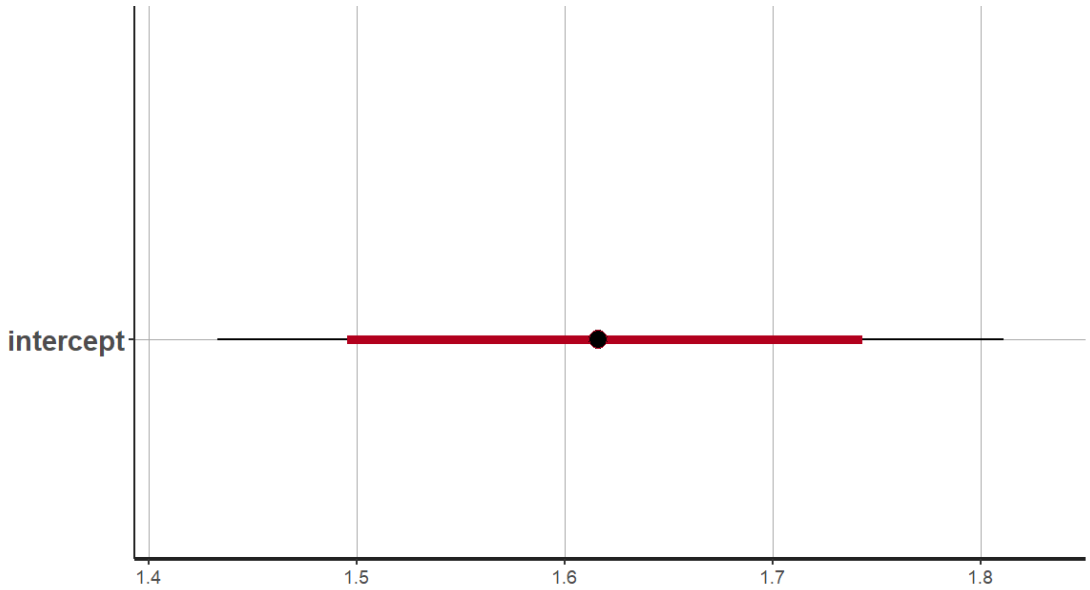


Figure 2: Variable Coefficients and their Respective Credible Interval

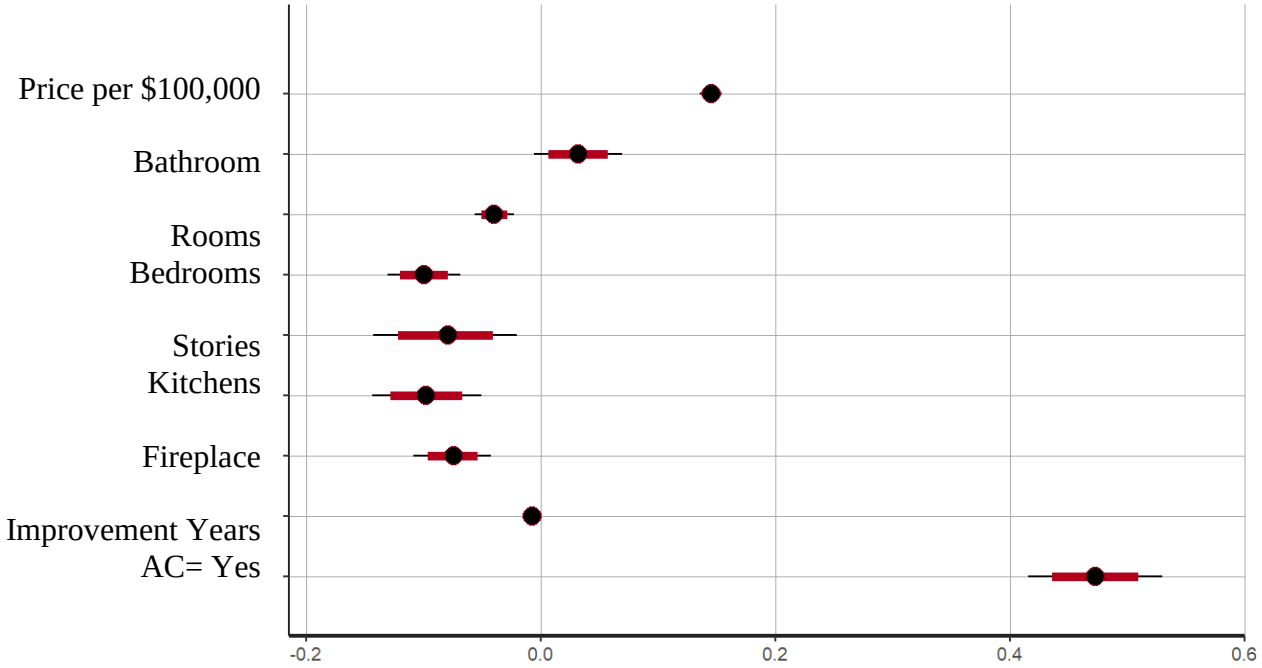
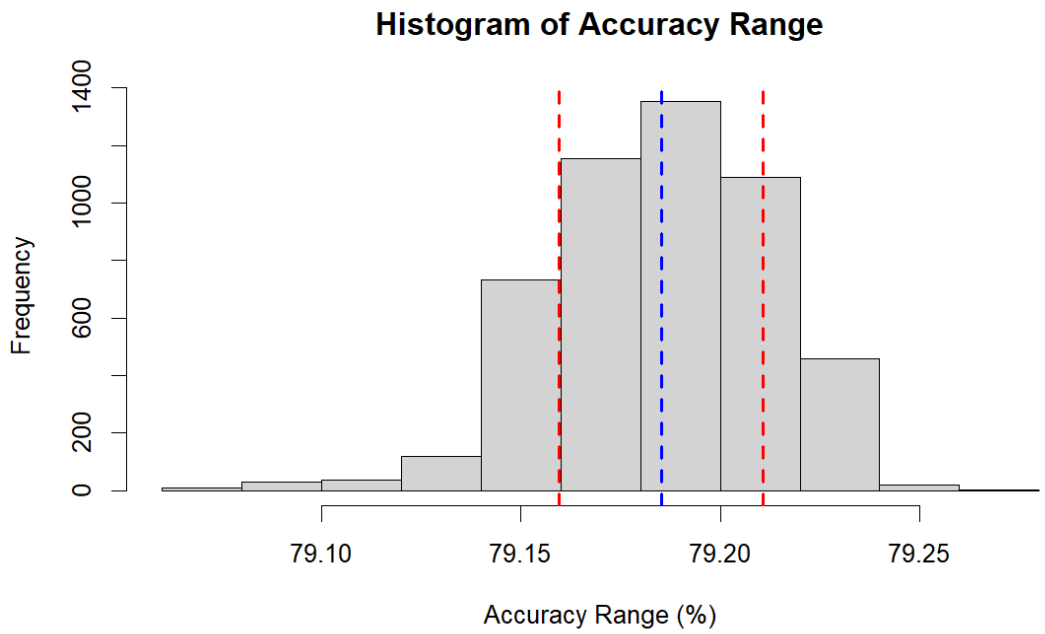


Figure 3: Accuracy Histogram for the Best Bayesian Logistic Regression Model



Footnote: The blue line is the mean of the accuracy; the red lines are the standard deviation of the accuracy.

References:

1. "Appendix A: An Introduction to Hierarchical Bayes Modeling in R." *Wiley Series in Probability and Statistics*, 13 Oct. 2006, pp. 279–322, https://doi.org/10.1002/0470863692.app1. Accessed 10 Sept. 2022.
2. Annis, Jeffrey, et al. "Bayesian Inference with Stan: A Tutorial on Adding Custom Distributions." *Behavior Research Methods*, vol. 49, no. 3, 10 June 2016, pp. 863–886, https://doi.org/10.3758/s13428-016-0746-9. Accessed 10 Apr. 2021.
3. "Bayesian Varying Effects Models in R and Stan." *Will Hipson*, willhipson.netlify.app/post/stan-random-slopes/varying_effects_stan/. Accessed 1 Apr. 2024.
4. "Hierarchical Models with RStan (Part 1)." *Biologyforfun*, 10 Nov. 2016, biologyforfun.wordpress.com/2016/11/10/hierarchical-models-with-rstan-part-1/. Accessed 1 Apr. 2024.
5. "Regression Models." *Mc-Stan.org*, mc-stan.org/docs/stan-users-guide/regression.html#multi-logit.section. Accessed 1 Apr. 2024.
6. Ren, You. "Bayesian Modeling of a High Resolution Housing Price Index." *Digital.lib.washington.edu*, 2015, digital.lib.washington.edu/researchworks/handle/1773/35321. Accessed 1 Apr. 2024.
7. Ren, You. "Bayesian Modeling of a High Resolution Housing Price Index." *Digital.lib.washington.edu*, 2015, digital.lib.washington.edu/researchworks/handle/1773/35321. Accessed 1 Apr. 2024.
8. Ram, Josh. "The Bayesian Housing Price for Competitions Is Finally Here — Part 1." *Medium*, 27 Dec. 2021, dronesai-peiskos.medium.com/the-bayesian-housing-price-for-competitions-is-finally-here-part-1-1c2dca097a0d. Accessed 1 Apr. 2024.
9. Garcia-Donato, Gonzalo, and Anabel Forte. "Bayesian Testing, Variable Selection and Model Averaging in Linear Models Using R with BayesVarSel." *The R Journal*, vol. 10, no. 1, 2018, p. 155, https://doi.org/10.32614/rj-2018-021. Accessed 11 May 2022.
10. Dogucu, A. A. J., Miles Q. Ott, Mine. (2021, December 1). *Chapter 13 Logistic Regression | Bayes Rules! An Introduction to Applied Bayesian Modeling*. Bayes Rules! An Introduction to Applied Bayesian Modeling; Bayes Rules. https://www.bayesrulesbook.com/chapter-13#chapter-13-prediction