Predicting Housing Price: A County Level Analysis

Group 16: Jieqi Tu (jt3098), Jiayi Shen (js5354)

Introduction

This study aimed to examine a collection of housing data, investigate the association between housing related factors, and present a final predictive model for housing price.

The dataset was downloaded from Kaggle.com. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. There are 21 columns in the original dataset with no NAs present (the description of each column can be found in Project Proposal). Variables “id”, “date”, “zipcode”, “latitude” and “longitude” were manually removed for simplicity. Also, “sqrt_living” and “sqft_lot” were manually deleted because of correlation with “sqrt_living15” and “sqrt_lot15”. In addition, variables “view” and “sqft_basement” and “yr_renovated” were converted to be binary for the purpose of easy interpretation. Moreover, because the dataset had more than 20000 observations, which would take a relatively long time to load and to run analysis, this project was only focused on a subset of the data (“view” > 0 and “bedroom” < 30). The cleaned dataset contained one response (price) and 12 predictors. Furthermore, because the original response vector was right skewed, Box-Cox transformation was conducted to ensure normal distribution of new response.

Exploratory data analysis/visualization

As shown in **Figure 1**, the distribution of price on its original scale was very right-skewed. The median price was $750000 whereas the mean was $939447. More than 6% of the houses were sold with > 2 million dollars, contributing to some outliers on the upper side. The scatter plots in **Figure 2** showed the individual relationship between 9 continuous and ordinal variables with price after Box-Cox transformation. Most houses sold in King County from 2014 to 2015 had condition of 3 to 5, and the range of grade was 4 to 13. It was not surprising to see that house sales price increased with increasing number of bathrooms/house, living room area, house area apart from basement and grade. The distribution of lot area was right-skewed and no clear trend could identified from the scatter plot. Year of construction did not seem to have an effect on its price.

Models

In modeling procedure for different types of regression, all the predictors were included. They were "bedrooms", "bathrooms", "floors", "waterfront", "condition", "grade", "sqft_above", "yr_built", "sqft_living15", "sqft_lot15", "basement" and "renovated".

In our study, both linear and non-linear models were included. For linear models, we fit least squares linear regression, ridge and lasso regression models, and compared them using root mean square error (RMSE) in the same resampled dataset. For non-linear regression models, we fit generalized additive model (GAM), multivariate adaptive regression splines (MARS).

Cross-validation (CV) was used to select the optimal tuning parameters. Cross-validation helps examine the prediction ability of the models obtained from the training dataset. In this study, 10-fold, 5 repeated cross-validation was used to select models. In each model, the same cross-validation tool (train function in caret package) was used in order to get comparable model performance. One of the advantages of cross-validation is that it does not require any assumptions for the model.

- Linear models (coefficients are shown in **Table 1**)

Using stepwise regression, no predictor was removed in least squares linear regression. In lasso regression, although it has the property of selecting variables, all predictors were keeped in lasso regression model. Moreover, the RMSE values for these three linear models are quite similar, but ridge and lasso regression had a lower range of RMSE than least squares regression. This might be due to the shrinkage of coefficients in lasso and ridge, which can reduce the predicted error.

- Non-linear models (coefficients are shown in **Table 2**)

We then tried Generalized Addition Model (GAM) (see **Figure 3**) and Multivariate Adaptive Regression Splines (MARS) to capture the non-linear relationship between individual predictors and the response, as well as interactions between variables. In our case, GAM outperformed MARS in terms of predictability (see **Figure 4**). This was possibly due to interactions between some predictors had significant influence on the response, and MARS accounted for those interaction terms.

MARS selected 14 out of 23 terms, and 9 of 12 predictors after tuning the number of degree and the number of terms. Grade, year built, and the area of living room were the most important whereas floors, bedrooms and renovated were not selected. Significant interactions selected by MARS were visualized in Figure xx. For example, houses with basements that had higher grades were more likely to be sold with a high price (**Figure 5**). It was unexpected that “grade” was the most important covariate selected by MARS whereas the scatter plot showed even distribution of price across years.

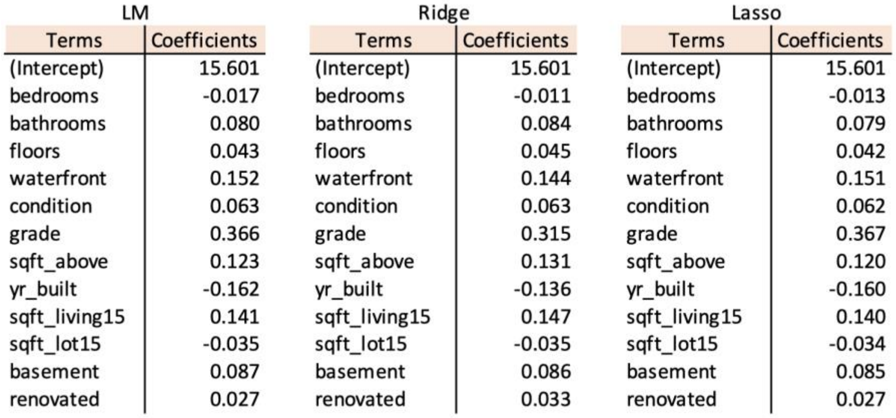
Conclusions

In general, nonlinear regression models were more likely to have lower RMSE values than linear models (**Figure 4**). This was because that MARS and GAM were capable of capturing nonlinear relationships between predictors and response. Therefore, non-linear models were more flexible than linear models. However, it is harder to make inference from non-linear models and make interpretations.

In terms of predictability, MARS showed better performance than other models discussed above. Grade, year of construction and a view to waterfront seemed to be the most important factors to consider for the house price in King County.

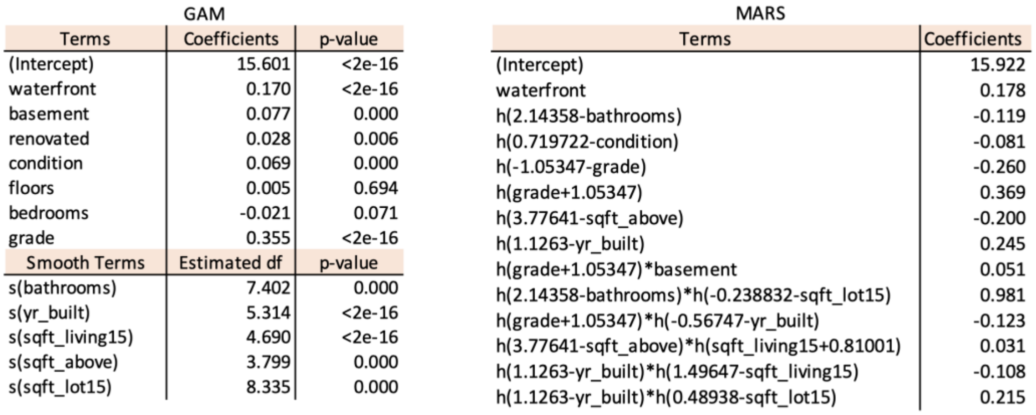
Appendix

Table 1. Coefficients for least squares regression, ridge regression and lasso regression



LM		Ridge		Lasso	
Terms	Coefficients	Terms	Coefficients	Terms	Coefficients
(Intercept)	15.601	(Intercept)	15.601	(Intercept)	15.601
bedrooms	-0.017	bedrooms	-0.011	bedrooms	-0.013
bathrooms	0.080	bathrooms	0.084	bathrooms	0.079
floors	0.043	floors	0.045	floors	0.042
waterfront	0.152	waterfront	0.144	waterfront	0.151
condition	0.063	condition	0.063	condition	0.062
grade	0.366	grade	0.315	grade	0.367
sqft_above	0.123	sqft_above	0.131	sqft_above	0.120
yr_built	-0.162	yr_built	-0.136	yr_built	-0.160
sqft_living15	0.141	sqft_living15	0.147	sqft_living15	0.140
sqft_lot15	-0.035	sqft_lot15	-0.035	sqft_lot15	-0.034
basement	0.087	basement	0.086	basement	0.085
renovated	0.027	renovated	0.033	renovated	0.027

Table 2. Coefficients for GAM and MARS



GAM

Terms	Coefficients	p-value
(Intercept)	15.601	<2e-16
waterfront	0.170	<2e-16
basement	0.077	0.000
renovated	0.028	0.006
condition	0.069	0.000
floors	0.005	0.694
bedrooms	-0.021	0.071
grade	0.355	<2e-16
Smooth Terms	Estimated df	p-value
s(bathrooms)	7.402	0.000
s(yr_built)	5.314	<2e-16
s(sqft_living15)	4.690	<2e-16
s(sqft_above)	3.799	0.000
s(sqft_lot15)	8.335	0.000

MARS

Terms	Coefficients
(Intercept)	15.922
waterfront	0.178
h(2.14358-bathrooms)	-0.119
h(0.719722-condition)	-0.081
h(-1.05347-grade)	-0.260
h(grade+1.05347)	0.369
h(3.77641-sqft_above)	-0.200
h(1.1263-yr_built)	0.245
h(grade+1.05347)*basement	0.051
h(2.14358-bathrooms)*h(-0.238832-sqft_lot15)	0.981
h(grade+1.05347)*h(-0.56747-yr_built)	-0.123
h(3.77641-sqft_above)*h(sqft_living15+0.81001)	0.031
h(1.1263-yr_built)*h(1.49647-sqft_living15)	-0.108
h(1.1263-yr_built)*h(0.48938-sqft_lot15)	0.215

Figure 1. Distribution of original response (left) and transformed response (right)

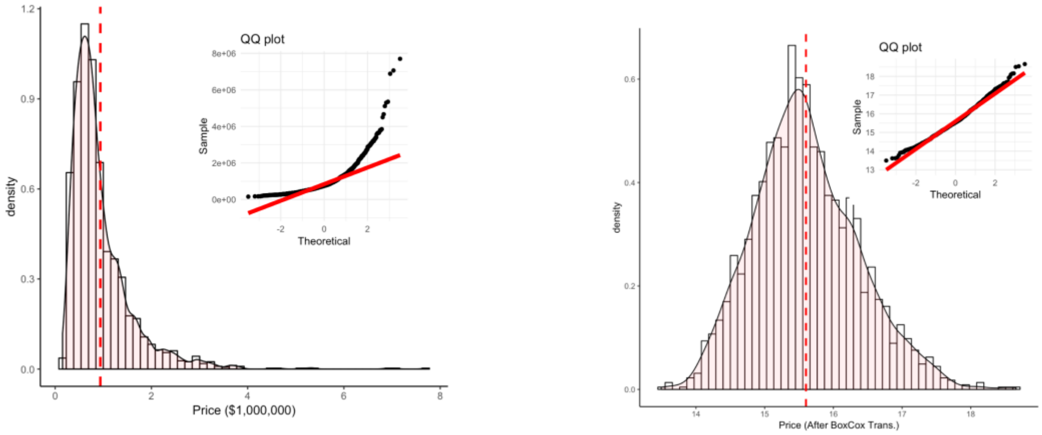


Figure 2. Scatter plots (left) of continuous variables and boxplots (right) of binary variables versus price

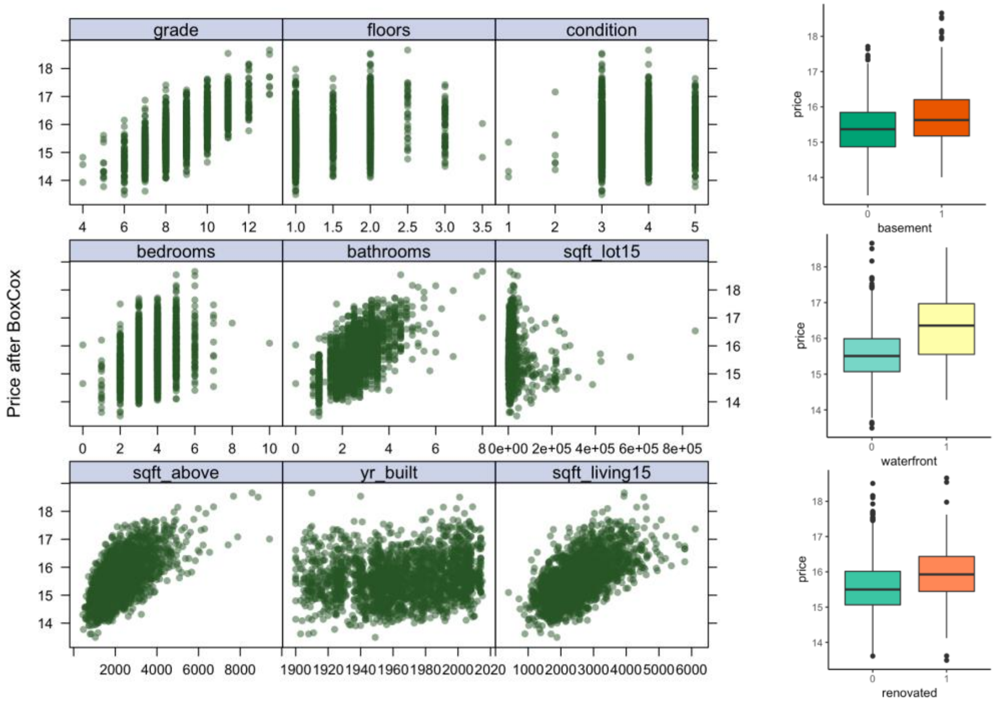


Figure 3. Partial dependence plots of GAM

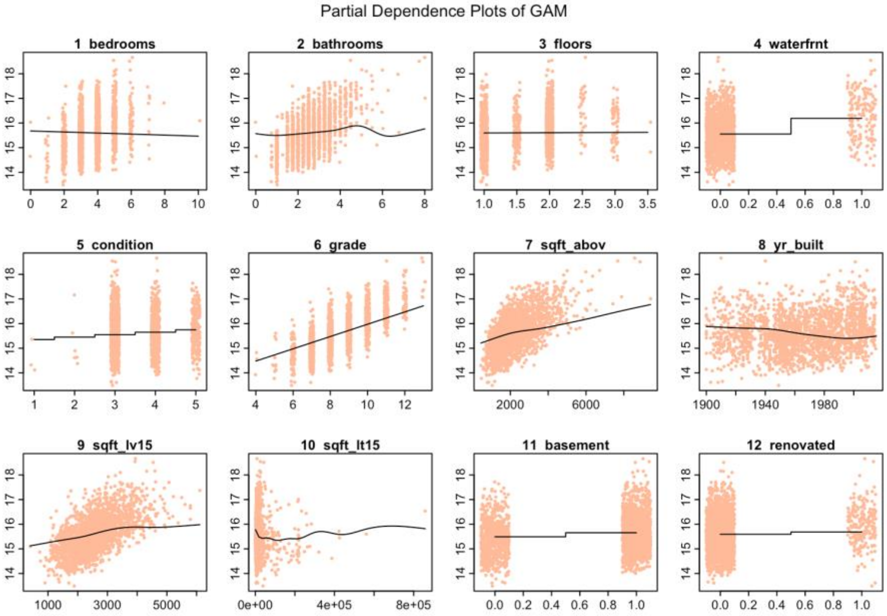


Figure 4. Boxplot of RMSE for ridge, lasso, least squares, GAM and MARS regressions

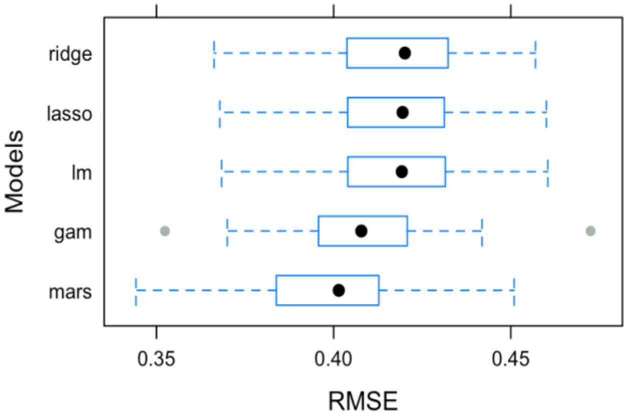
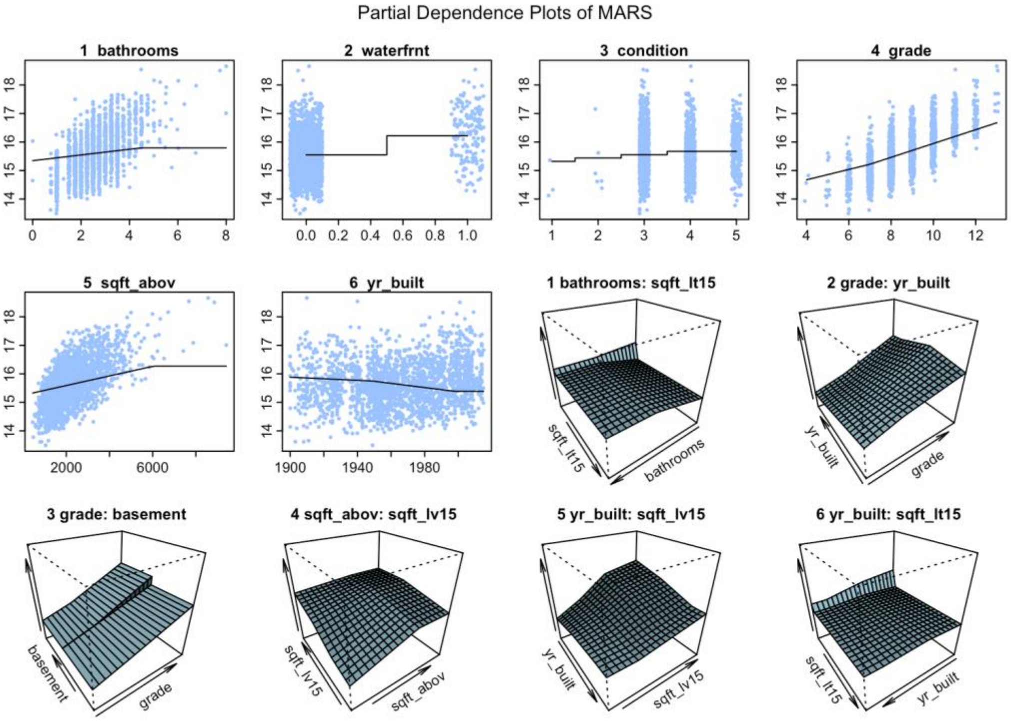


Figure 5. Partial dependence plots of MARS


Partial Dependence Plots of MARS
1 bathrooms
2 waterfrnt
3 condition
4 grade
5 sqft_abov
6 yr_built
1 bathrooms: sqft_lt15
2 grade: yr_built
3 grade: basement
4 sqft_abov: sqft_lv15
5 yr_built: sqft_lv15
6 yr_built: sqft_lt15
sqft_lt15
bathrooms
yr_built
grade
basement
grade
sqft_lv15
sqft_abov
yr_built
sqft_lv15
sqft_lt15
yr_built