Big Data Project 1

Machine Learning on House Price

Jiawang Zhou, Xi Liu

03/03/2019

1. Background

As both people's daily residence and investment target, houses have always been an important topic to study. This report analyzes the factors that decide house price in Ames, Lowa and makes prediction with the glm model. Of the 81 variables in the train dataset, we picked 18 for data description. The dependent variable is the sale price of houses and independent variables include features such as lot size, remodel year, number of bathrooms, living area above ground, overall quality and so on. We first used descriptive statistics to know about our data. After data processing, we used knn, cart, svm, ridge, lasso, and glm methods to build models separately and chose the best one, the glm model with highest adjusted R^2 and least MAE and RMSE. Finally, based on the glm model, we made prediction of house price with test data.

2. Data Description

1) Variables Overlook

Our data was obtained from Kaggle, including a train dataset with sale price and a test dataset without sale price. The train dataset includes 1460 observations and 81 variables and the test dataset includes 1459 observations and 80 variables.

2) Variables Selection

The 18 variables below were chosen based on logical judgement and the correlation between sale price (Shown in Appendix 1). Then we tried to find variable importance with a quick Random Forest (Shown in Appendix 2), and we wanted to get an overview of the most important variables including the categorical variables.

The 18 variables include:

Variable Name	Variable Type	Variable Description
SalePrice	Numerical	Sale price of the house
LotFrontage	Numerical	Linear feet of street connected to property
LotArea	Numerical	Lot size in square feet
YearRemodAdd	Numerical	Remodel date (same as construction date if no remodeling or additions)
MasVnrArea	Numerical	Masonry veneer area in square feet
BsmtQual	Categorical	Evaluates the height of the basement
TotalBsmtSF	Numerical	Total square feet of basement area
Fireplaces	Numerical	Number of fireplaces
CentralAir	Categorical	Central air conditioning (Yes/No)
GrLivArea	Numerical	Above grade (ground) living area square feet
FullBath	Numerical	Number of full bathrooms

HalfBath	Numerical	Number of half bathrooms
BedroomAbvGr	Numerical	Bedrooms above grade (does NOT include basement bedrooms)
KitchenAbvGr	Numerical	Kitchens above grade
GarageArea	Numerical	Size of garage in square feet
PavedDrive	Categorical	Paved driveway (Paved, Partial Pavement, Dirt/Gravel)
Neighborhood	Factor	Physical locations within Ames city limits
OverallQual	Categorical	the overall material and finish of the house (1 Very poor – 10 Excellent)

Table 1: Variable Description

3) Data Description

Firstly, we tried to find the distribution pattern of our dependent variables: Sale Price. We can tell that in our case, the distribution of log-form sale price is similar to normal distribution pattern. (Show in Appendix 3)

Then, we calculated the correlation between all the variables. The first five variables with highest correlation with sale price are: overall quality, living area above ground, garage size, total basement area and number of fullbath. (Shown in Appendix 1) The highest correlation is 0.79, between overall quality and sale price, indicating overall quality is most influential on house price in this place. There is a significant position relationship between overall quality and sale price. The price of houses with quality level 1 and 2 is similar. But as quality increases from 3 to 10, the price difference between each level becomes increasingly larger. (Shown in Appendix 4) Also, as above-ground living area increases, house price also tends to increase. (Shown in Appendix 5)

Also, we found that sale price is closely affected by the location of the house. The boxplot is shown below. In several certain neighborhoods, the price is obviously higher than others, for example, NridgHt, NoRidge and StoneBr. Additionally, in neighborhoods with most houses on sale, the price is usually not high. (Shown in Appendix 6 and 7)

3. Data Processing

The original dataset includes both train and test data. We first combined the two datasets into one and processed the raw data. Then we divided the total dataset back to train and test datasets. The data processing methods include:

1) Merge Relevant Variables

The original dataset includes the number of full/half bathrooms in the basement and above ground. We summed all the full bathrooms and half bathrooms in the house and got variables FullBath and HalfBath.

2) Dependent Variable Logarithm Transformation

The dependent variable in our model is the sales price. Its original data is right-skewed with a skewness of 1.88. (Shown in Appendix 3) We used its log form to get an approximately normal distribution. Similarly, for all other independent variables with skewness above 0.75 in the dataset, we used their log form in the model.

3) Process NA numbers

There are lots of NAs in the Kaggle dataset, which will influence our analysis. For all numerical variables with a missing value, we used the mean of that feature to replace NA. For all categorical variables with a missing value, we use 0.

4) Process Categorical Data

In order to fit categorial data into machine learning models, we use caret dummyVars function for hot one encoding for categorical features, which means we change categorial features into many variables with 0 or 1.

5) Outliners

There are several outliners in the dataset. We have deleted row 524 and 1299 in order to reduce bias of our dataset.

4. Model

1) Model Training

We used resampling methods named repeated cross-validation to create a set of modified datasets from the training samples. Then we applied knn, cart, svm, ridge, lasso and glm methods separately to train the model. The algorithm selection standards include adjusted R^2, MAE and RMSE. The accuracy of all the algorithms is shown in the chart below. The result shows that the glm algorithm provides largest adjusted R^2 and smallest MAE and RMSE. Therefore, it is the most accurate and effective method to build model and make prediction.

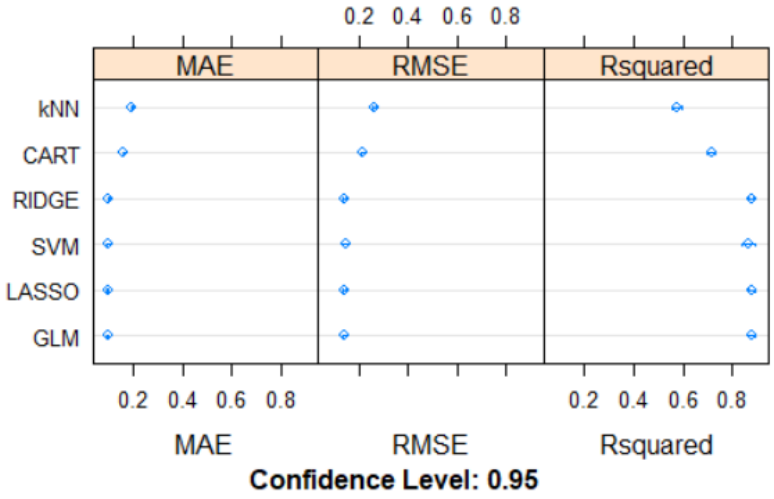


Chart 1: Accuracy of All Algorithms

2) Model Validation

We put the predicted sale price and actual sale price into a scatter plot to reveal the validation

of our prediction. The red line is a 45° line, indicating y=x. We can see that the prediction result is quite accurate, with only few biased points. This result shows that our model has a strong prediction accuracy.

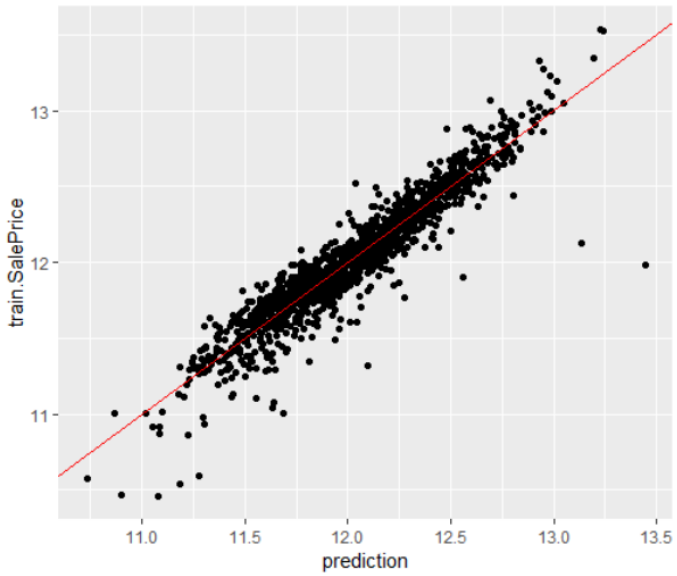


Chart 2: Prediction Result

3) Model Prediction

Then we tried to put our model into the test data set. We made prediction based on the available variables. The distribution of our prediction is also pretty similar to the normal distribution according to our Q-Q plot. (Shown in Appendix 8)

5. Conclusion

In conclusion, the GLM model is the best fit to make house sale price prediction by providing the highest adjusted R^2. The result amazed us a little bit since usually the Non-Regression model should achieve higher accuracy. Maybe we need to try out more machine learning models on this problem or to develop and suitable model ourselves in the future.

6. Improvements

There are several improvements that we may promote our machine learning model further.

1) Variable Selection

There are over 81 variables in our dataset, for simplicity we just selected 18 of them. The selection itself has eliminate most part of data which could be useful in machine learning method.

2) Outlier removement

For the time being, we were keeping it simple by removing the only two really big houses

with low Sale Price manually. However, we intend to investigate this more thorough in a later stage (possibly using the 'outliers' package).

3) Model Selection

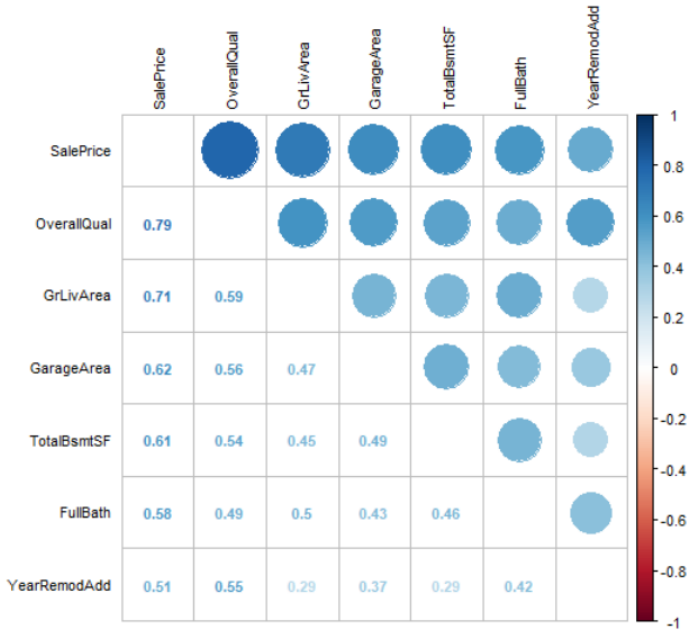
Firstly, though we use the R square and RMSE for accuracy estimation, but there are other measurements of model accuracy, under which GLM might not be the best model.

Secondly, we have used the trial and error method with many different algorithms, unfortunately, there is too many machine learning algorithms out there that may be the better fit for this dataset.

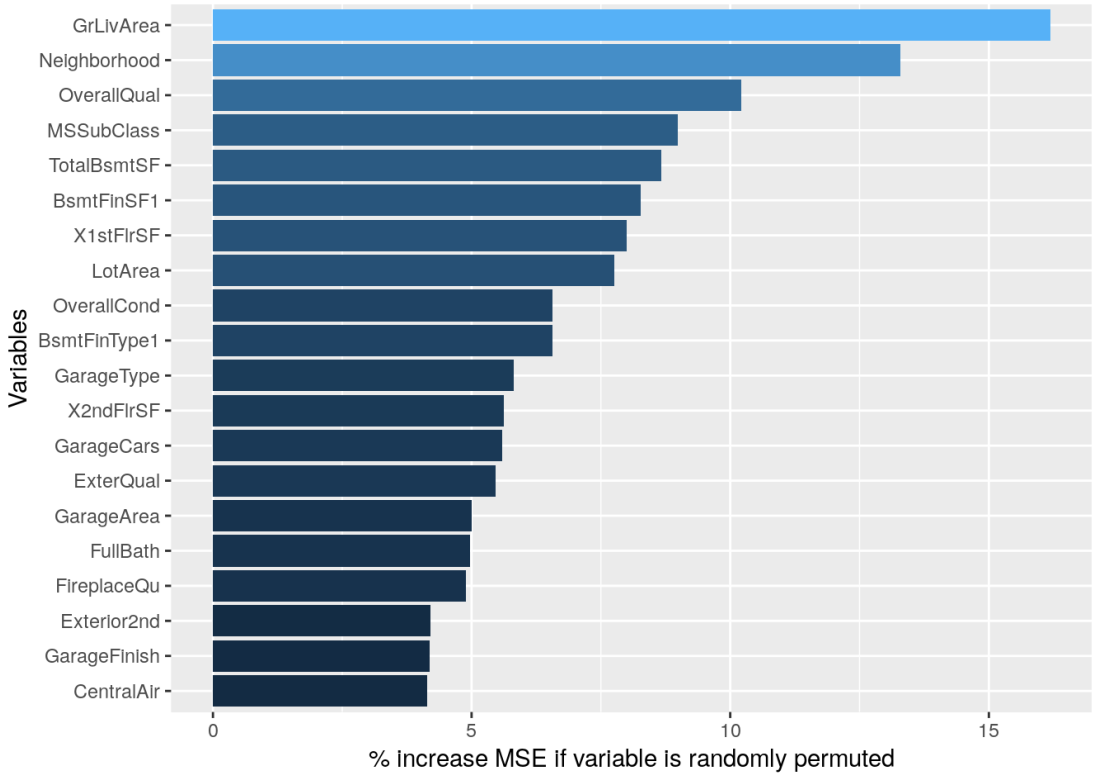
Lastly, instead of using a single glm model for testing, many machine learning papers are using mix models with weighted averaging prediction.

Appendix

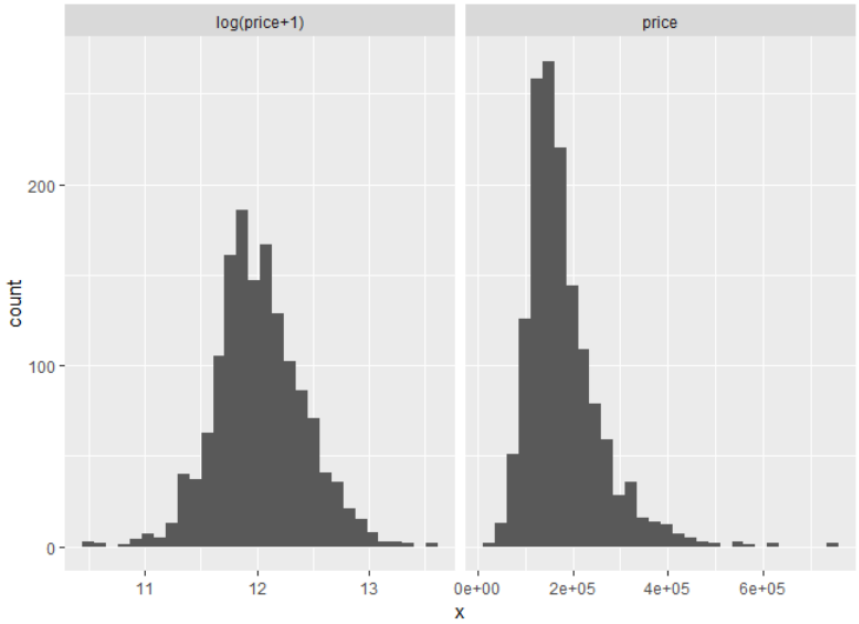
Appendix 1: Correlation Plot of Variables



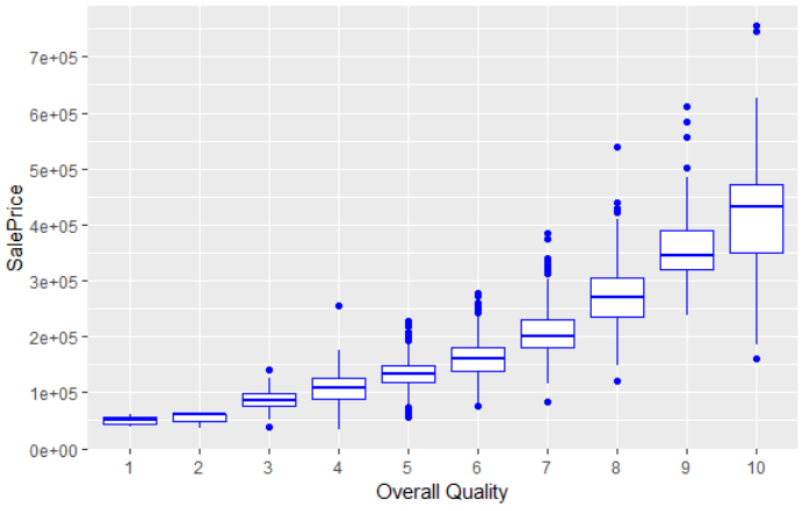
Appendix 2: Quick Random Forest Variable Importance



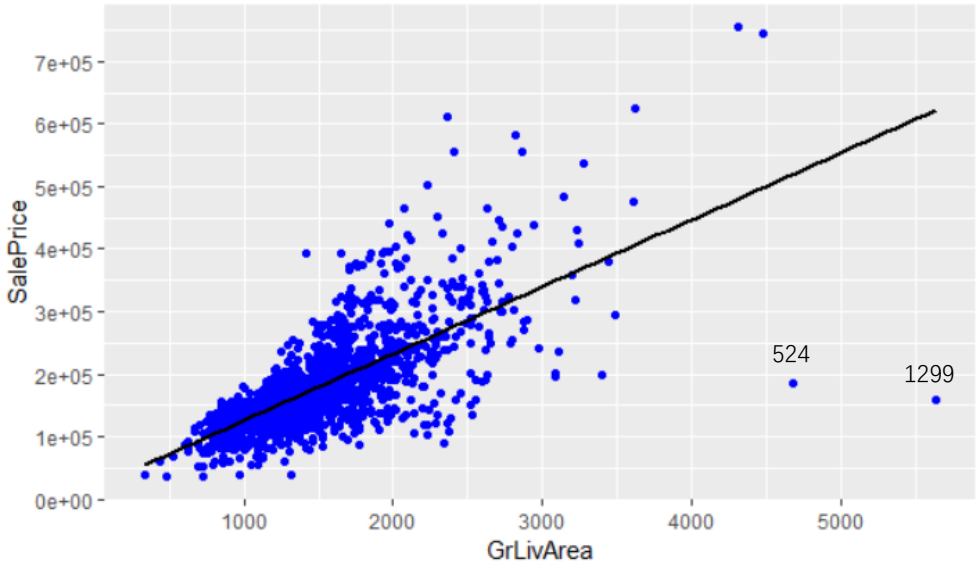
Appendix 3: Data Description & Processing – Sale Price



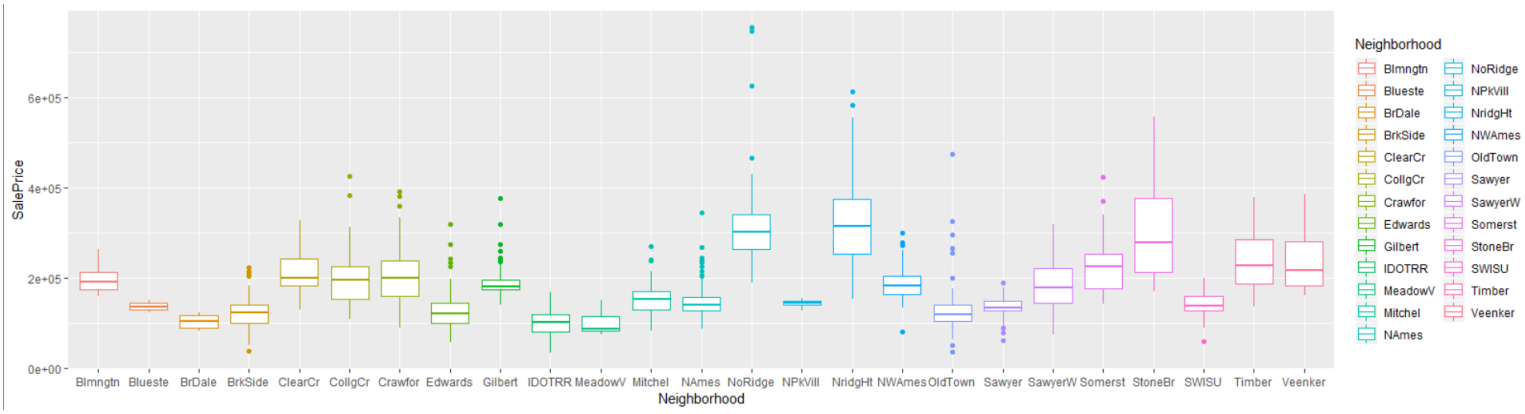
Appendix 4: Relationship between Overall Quality and Sale Price



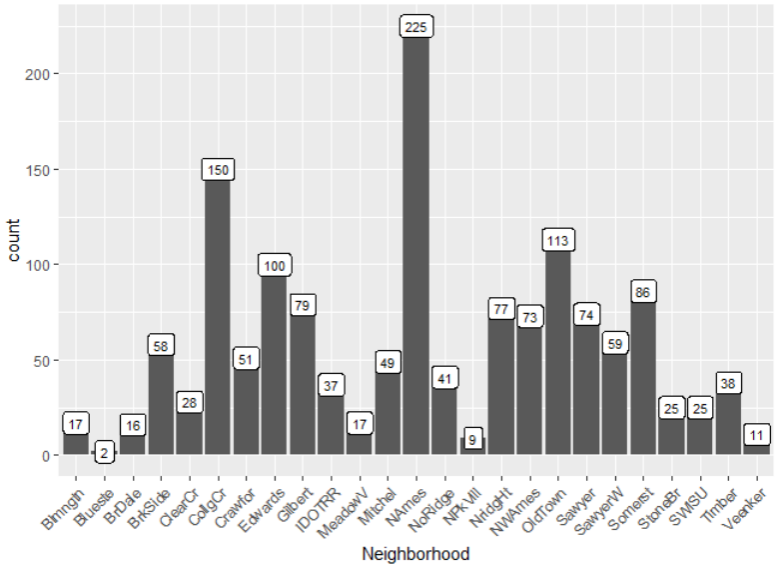
Appendix 5: Relationship between living area above ground with sale price



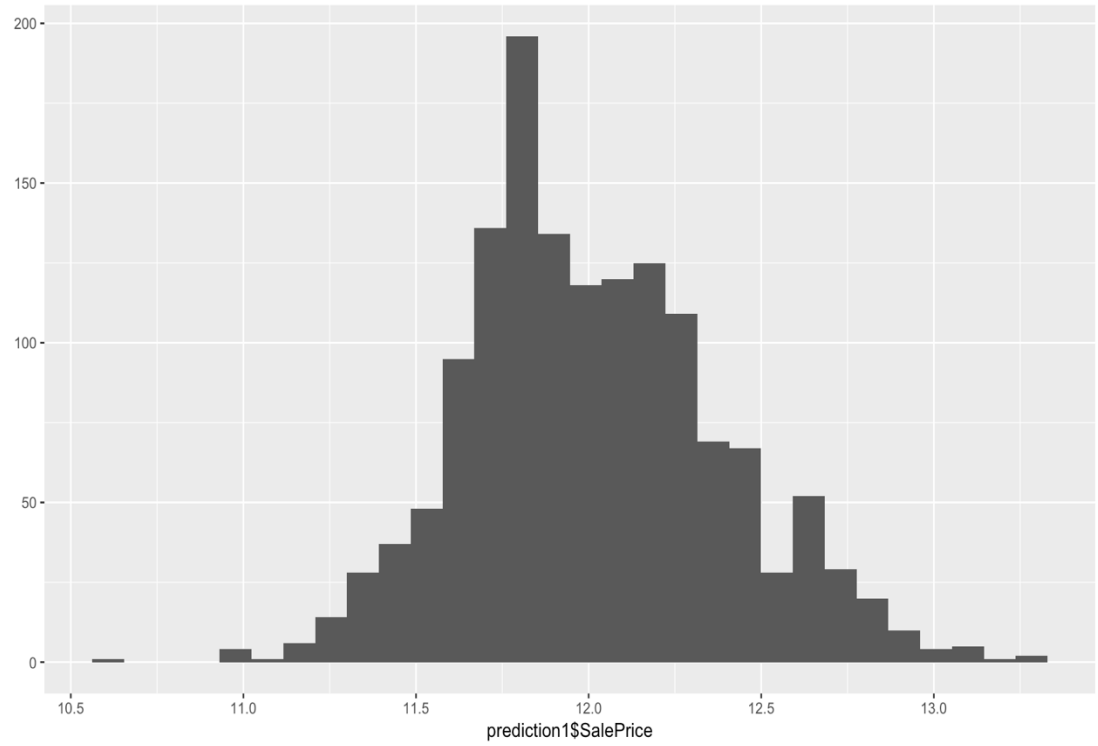
Appendix 6: Boxplot of Neighborhood and Sale Price



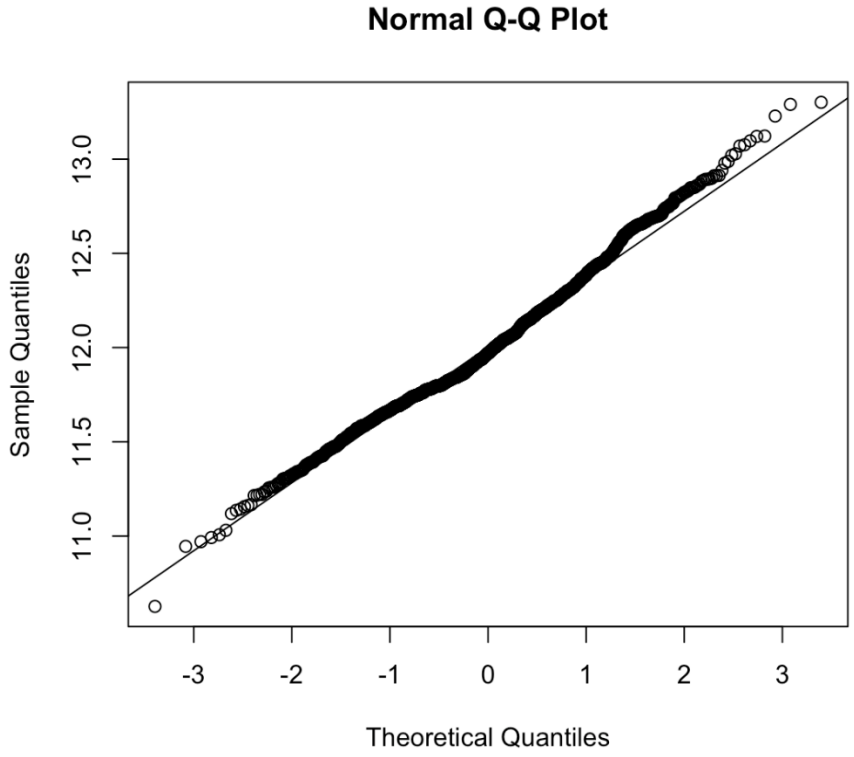
Appendix 7: Distribution of Houses in Different Neighborhoods



Appendix 8: Prediction Result distribution



Appendix 9: Q-Q Plot of Model


Normal Q-Q Plot
Sample Quantiles
11.0
11.5
12.0
12.5
13.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles