House Price Regression, Regression Report

Ashok Kamath and Daeyeop Kim[1]*

Abstract
In this project, we want to predict the sale prices of housing.

Keywords
House — Prices — Prediction

[1] *Computer Science, School of Informatics , Computing and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

For the housing prices dataset, the problem is predicting the sale prices of the houses and the objective is to use the attributes in the dataset to predict the sale price.

For this dataset, the training set has 1460 rows and 81 columns while the test set has 1459 rows and 80 columns. Therefore, the train-test-split ratio is about 50-50.

The columns describe features of each house such as the Overall Quality, the neighborhood, land slope, year built, roof style, bedroom, kitchen, and gross living area. Since there are so many columns, it could be productive to find the columns that are most correlated with Sale Price and use those in regression.

To be more specific about the meaning of some ambiguous columns, Overall Quality refers to the overall finish and material quality while Basement Exposure refers to whether the basement is walkout or garden level.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

Since the number of columns was so high, we decided to first find the columns that were the most correlated with sale price and only work with those columns for prediction. For that reason, we only need to handle the missing values in that subset of columns.

After finding the columns with the greatest correlation to Sale Price, we had only two attributes that had missing values, Garage Year Built and Masonry Veneer Area, which had 81 and 8 missing values respectively. From there, we looked at the distribution of both attributes to choose the most reasonable method of filling the missing values. For Garage Year Built, the average and the median are about the same, so either would work and we chose to use the mean. For Masonry Veneer Area, the median was 0 while the mean was 103, indicating a right skewed distribution, so it made sense to fill the missing values with the median, 0.

2.2 Exploratory Data Analysis

To further analyze the relationships between Sale Price and the subset of columns that had high correlation with Sale Price, we chose to create scatter plots.

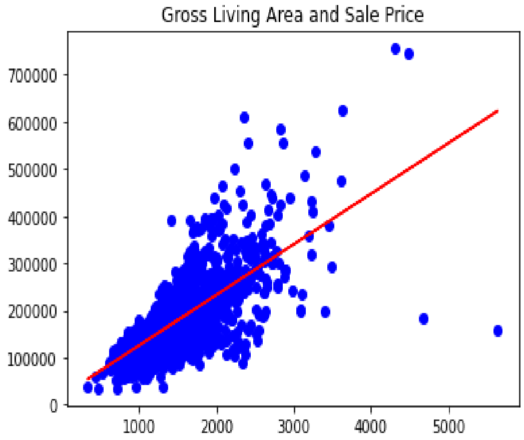

Figure 1. Scatter plot of Gross Living Area with Sale Price

The first scatter plot shows that Gross Living Area and Sale Price have a strong positive relationship, but there are some outliers with high gross living area yet they have a low sale price.

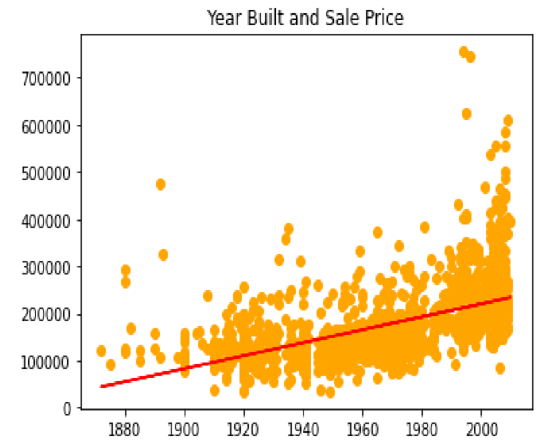

Figure 2. Scatter plot of Gross Living Area with Sale Price

The second scatter plot, which shows the relationship between Year Built and Sale Price, indicates a slight positive relationship between the two variables.

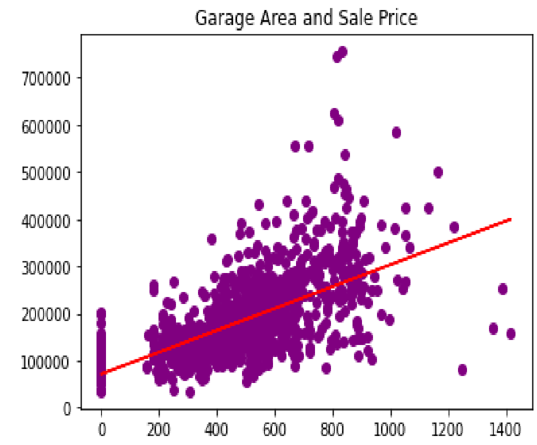

Figure 3. Scatter plot of Garage Area with Sale Price

The third scatter plot shows that Garage Area and Sale Price have a stronger positive relationship than Year Built but not as strong as between Gross Living Area and Sale Price.

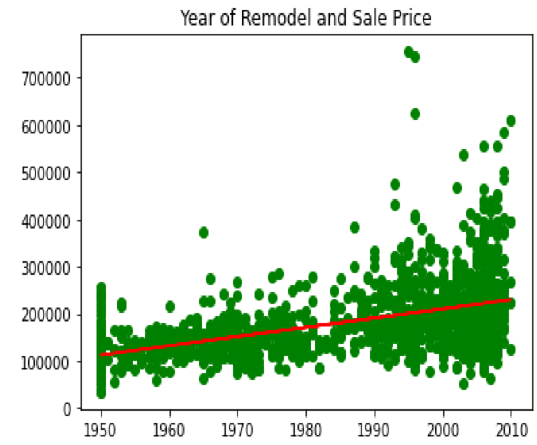

Figure 4. Scatter plot of Year of Remodeling with Sale Price

The Year of Remodeling has the weakest positive relationship with Sale Price of the 4 scatter plots.

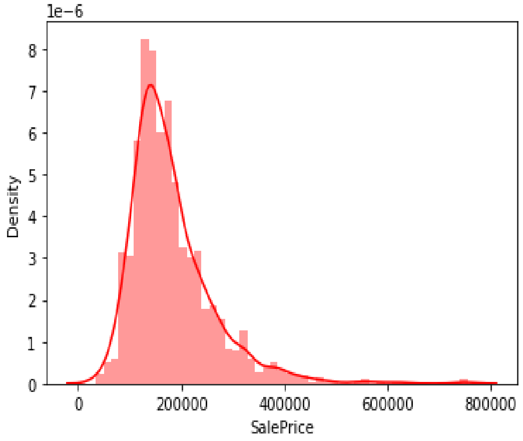

Figure 5. Sales price distribution

We also wanted to see the distribution of some of the attributes that were included in the subset of the data we selected earlier. For Sale Price, we found the data is mostly normally distributed with a slight right skew and gross living area had a similar distribution, which would likely explain the strong correlation between the two variables.

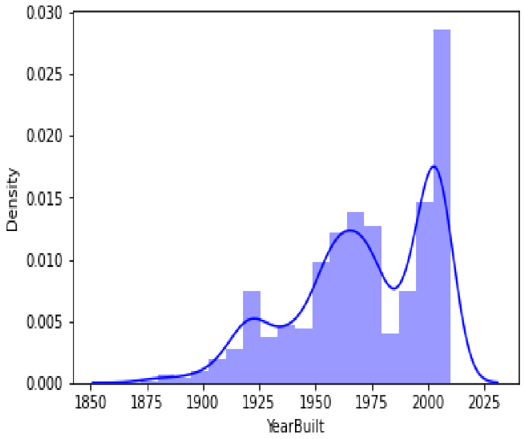

Figure 6. Year Built distribution

Year Built had a distribution that was left skewed. There were very few houses that were built in the 19th century and most were built after 1950. There were many houses that were built after 2000. The latest a house was built was 2010 and the earliest was 1872.

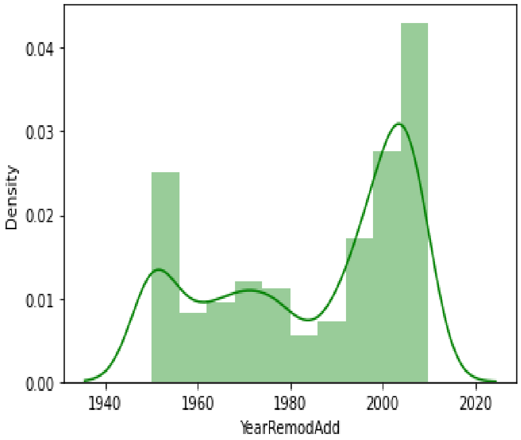

Figure 7. Year Remodeling distribution

The Year of the Remodeling tended to be either quite recent or quite old, as far back as earlier than 1960. The mean year of remodeling was 1985 while the mean was 1994.

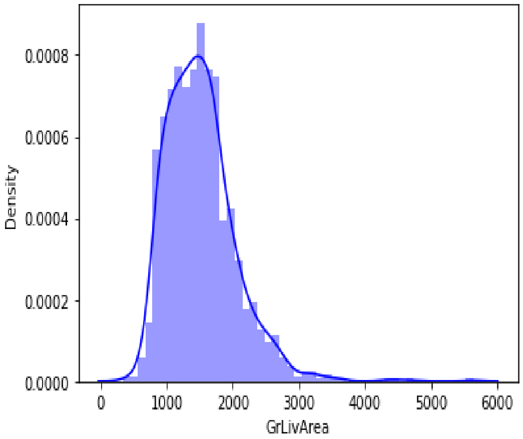

Figure 8. Gross Living Area distribution

The Gross Living Area has a distribution that is right skewed. There were lots of data points from 500 to 3000 .

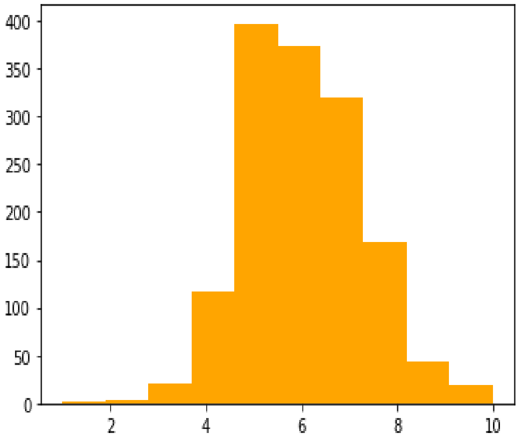

Figure 9. Overall Quality distribution

Overall Quality was normally distributed with the median and mean at 6.

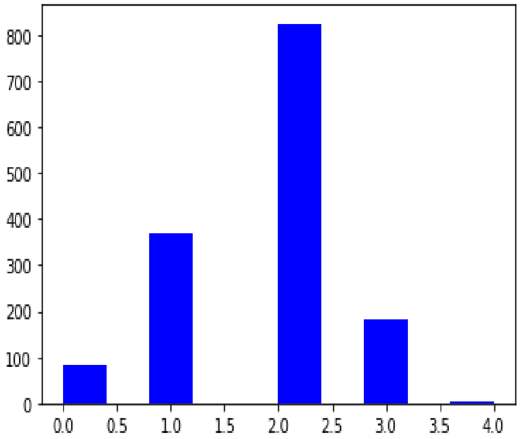

Figure 10. Garage Cars distribution

For Garage Cars, the data was normally distributed but with a slight left skew since the mean was slightly less than the mean. The maximum number of Garage Cars was 4 while some houses had no Garage Cars.

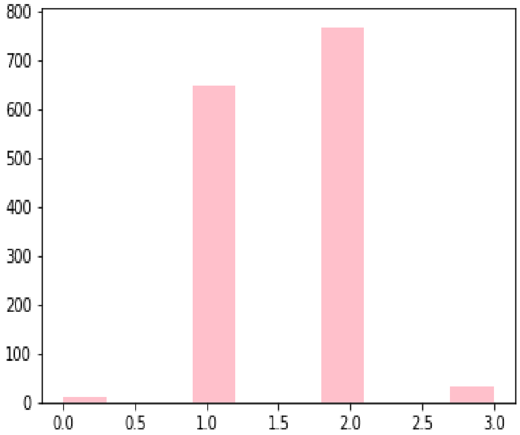

Figure 11. Full Bath distribution

For Full Bath, at least 50 percent of houses had at least 2 while some had none and the maximum number of full baths was 3.

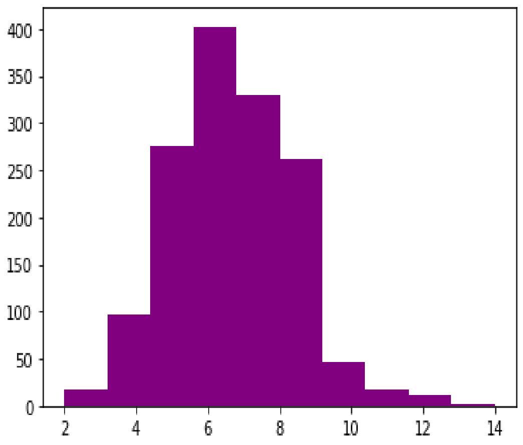

Figure 12. Total Rooms Above Ground distribution

Additionally, Total Rooms Above Ground was normally distributed with the maximum being 14 and the minimum being 2. The data was slightly right skewed since the mean was 6.5 and the media was 6.

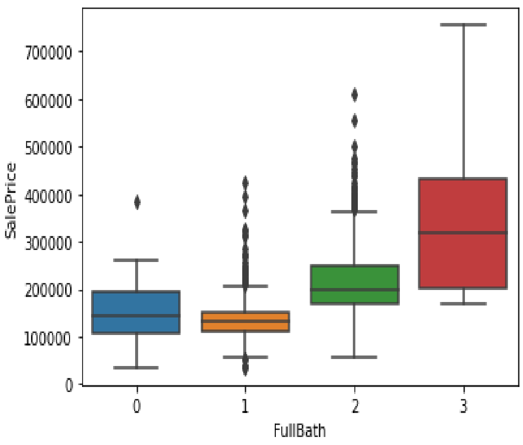

Figure 13. Boxplots of Full Bath with Sales price

Furthermore, our boxplots indicate that with more Full Baths, the price of the house likely increases. However, there is a lot of variance when the house has 3 baths and there are outliers that could affect regression results when there are 0, 1, or 2 baths.

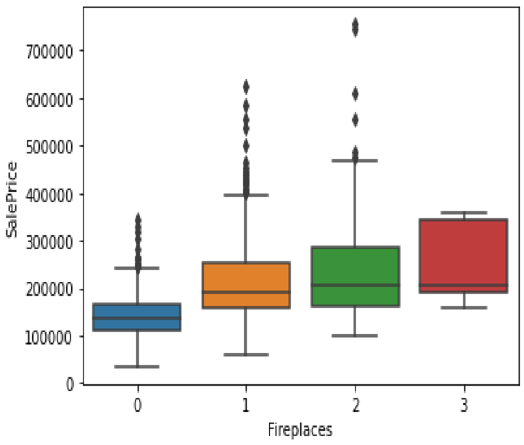

Figure 14. Boxplots of Fireplace with Sales price

For Fireplaces, on the other hand, when there are 3, there is not much variance in the sale price, but there is a strong right skew. For 0, 1 and 2 Fireplaces, the data is more evenly distributed, but there are more outliers.

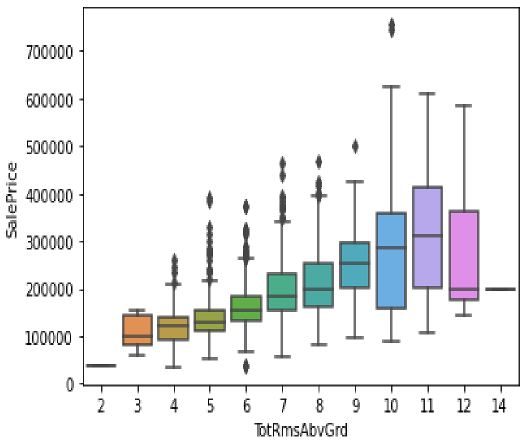

Figure 15. Boxplots of Total Rooms Above Ground with Sales price)

For the Total Rooms Above Ground attribute, there is hardly any variance in Sale Price when there are only 2 or 14 rooms above ground, which is probably due to a lack of entries with this number of rooms above ground. The boxplots indicate that as the total rooms above ground increases, the Sale Price typically will increase. For rooms above ground ranging from 4-10, there are outliers on the positive side.

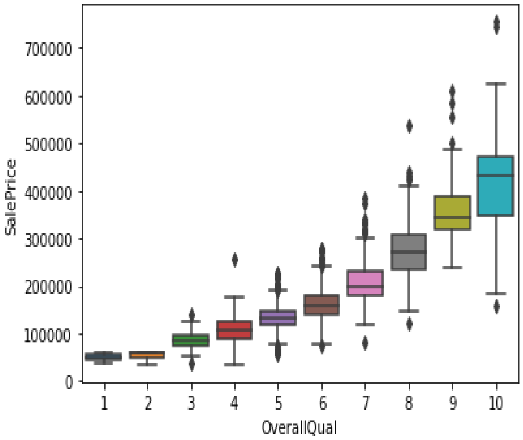

Figure 16. Boxplots of Overall Quality with Sales price)

Finally, Overall Quality has a strong positive relationship with Sale Price. There is little variance in price when the Overall Quality is 1,2, or 3, but the variance in price increases as the Overall Quality increases.

3. Algorithm and Methodology

For predicting the sale prices of housing, we used linear regression to start. This algorithm works by minimizing the sum of squared errors when creating a line of best fit through the training data. The line of best fit for the training data is then used to predict the appropriate values for the entries in the testing data.

In addition to linear regression, we used a random forest regressor, which is similar to the random forest classifier that we used on the spaceship titanic dataset. The random forest regressor, however, averages the results of the decision trees in the forest rather than take the majority vote. We also used a support vector regressor which attempts to predict new entries by first projecting the data into a higher dimension.

Besides linear regression, random forest and support vectors, we also used the gradient boosting regressor and a simple decision tree regressor. The gradient boosting regressor continually improves a decision tree model by calculating the gradient on the errors of the previous decision tree model. The decision tree regressor, as opposed to classifier, outputs a prediction by using the average of the training data in the leaf node that the decision nodes lead to based on the test entry.

Furthermore, we used Bayesian Ridge Regression, which employs Bayesian methods to make up for insufficient or weakly distributed data. In tandem, we used Lasso Regression, which performs a shrinkage technique that makes the data values shrink towards a central point. Lastly, similar to the classifier models for spaceship titanic, we created an MLP and KNN regressor.

4. Experiments and Results

Accuracy percentage with different model

Threshold percentage with 9 Models	
Model Name	Percentage threshold
Linear Regression	71.8187%
Random Forest Regression	81.9776%
Support Vector Regressor	-5.0832%
Gradient Boosting Regressor	84.2915%
Decision Tree Regressor	67.4467%
Bayesian Ridge Regression	72.1804%
Lasso Regression	71.8188%
MLP Regressor	68.2111%
KNN Regressor	72.1854%

Out of all the models, the Gradient Boosting Regressor model performed the best with a cross validation accuracy rate of about 84.29%. We were impressed with this rate of accuracy since most of the other models were not able to pass the 80% threshold. This demonstrates why the Gradient Boosting algorithm has been so popular within the data science community lately.

The only other model that was able to pass the 80% threshold was the random forest regressor model, which for cross validation on the training data had an accuracy rate of about 82%. This was another model that has its foundations in decision trees, except random forest is not as strategic in improving performance with the creation of each new decision tree, which could explain why it was not as accurate as the Gradient Boosting Regressor.

The KNN Regressor and the Bayesian Ridge Regressor had accuracies slightly above 72% while Lasso and Linear Regression models had accuracies that were slightly below 72%. The MLP Regressor had an accuracy of about 69% for the training data in cross validation while the decision tree model had an accuracy of about 64%. Lastly, the support vector regressor in cross validation failed to report a reasonable accuracy rate.

Since the Gradient Boosting Regressor performed the best in cross validation, we used that model on the testing set and submitted our results to Kaggle, which reported we had an error of .15 while the best of all time submission had an error of .00. We placed on the leaderboard at 2474 out of 4130 spots on the leaderboard.

5. Summary and Conclusions

For this section, the goal was to predict the sale prices of housing. The dataset had 80 attributes, excluding the target variable of sale price. All of the features described the house that was sold and some examples include Overall Quality, year built, roof style, number of bedrooms and gross living area. For handling the missing values in this dataset, we employed the same reasoning as we did for the Spaceship Titanic dataset.

In our exploratory data analysis, we only worked with columns that had the greatest correlation with sale price since the number of attributes was high, so we wanted to narrow it down. We noticed a strong positive relationship between Gross Living Area and Sale Price via a scatter plot. Additionally, we found Sale Price was normally distributed but with a slight right skew, similar to the distribution for Gross Living Area. Our boxplots showed that a house with more Full Baths will likely have a higher price, but there is more variance as the number of full baths increase. There was a similar relationship between Overall Quality and Sale Price since the variance increased as the Overall Quality increased.

To predict the sale prices of housing, we used various models ranging from Bayesian Ridge and Lasso Regression to KNN Regression. Through cross validation on the training data, we found that the Gradient Boosting Regressor performed best so we used that on the test data, giving us an error rate of .15, which had us place at 2474 out of 4130 spots on the leaderboard while the leading submission, in rank 1, had an error rate of .00. Our model could be improved not only by performing a grid search for each model, but also removing outliers, using columns besides those that are highly correlated with sale price, and other techniques.

References