Predicting Housing Prices

*Predictive Analytics Project*

Jingru Zhang

Bellevue University

**Abstract**

Due to the mortgage interest rates falling recently, there are more and more property transactions occurring. And the house is one of the largest and most expensive purchase most people made in their lifetimes, thus, I can imagine most people would not do a quick decision to purchase a house, instead they will take many factors into consideration, for instance, location, size, price, and so on. In addition, house is a big business, the total real estate market is around 16 trillion in United States, tons of people work as house business agents, it is also important for these people to know the worth of houses. The purpose of my project is to use real world house information to predict the house price. Zillow is one of the largest and most trusted marketplaces for houses, it also provides relatively rich house information, therefore, I decided to collect my own dataset by scrapping the house information from Zillow.com. The dataset includes information including the number of bedrooms, home type, year built, school rating and so on. All the houses scraped are in the Omaha metro area. In order to achieve the desired predictive power, I need to clean the data, remove or fix the outliers and miss values etc. Then I analyzed the correlation between independent variables and dependent variable. After that I used the proper variables to train the models. I used several algorithms to train the models, including linear regression, support vector machine and random forest. Then I used the test data to see the model performances on test data and see how well the models trained on training data can be generalized to the test data.

**Background**

A house is one of the most valuable economic assets a person can purchase during his/her life. Thus, it is very important to make the right decision whether to buy a house or not and at the right price as well. In common, different people might have different reasons to buy houses, for example, a family has a school age kid or school age kids might be more interested in buying a house which has a good public school in the region, people do not have a big budget might look for a relatively cheap house. Also, some houses are pretty old yet still have high listed prices for sale. I also have heard of something like in some places, the size of garage is highly related to the house price, and there are also clichés like “location, location, location!” What are the important factors to determine the price of a house, can we accurately predict the house price with the house information? If the model is able to predict the house prices, it could be used by those people who are looking for a new house, the potential buyer could compare the listed price with the predicted price and see whether it is a good invest or not. It could also be helpful for real estate agents, who can use the model to provide a reasonable listed price for a house for sale to meet the customers’ need and possibly maximize the profit out of it.

**Methods**

The CRISP-DM methodology was used to analyze the data. It starts with business understanding, which focuses on understanding the project objectives and requirements. The next stage of the CRISP-DM process is data understanding, including data collection, exploring data and identifying data quality. The third stage is data preparation, cleaning data, generating new feature if necessary, integrating data. The next stage is data modeling, namely, choose and apply the appropriate modeling techniques. Next, I evaluate model performance on the test data.

For business understanding, as I mentioned before, the motivation and goal of this project is to predict housing prices which can not only help house buyers to make a good decision but also be beneficial for house business agents to have a better idea about the worth of houses.

There are some public housing prices dataset available, but the house price changes over time, thus, I decided to scrape the most recent data directly from Zillow.com. Given a city name, for instance, Lincoln NE, the Zillow website shows a set of house cards with basic information, such as price, size, bedroom, bathroom, etc. And if one further clicks a card, it will bring one to the website reserved for the specific house and more information will show up, for instance, year built, garage, cooling, roof type, school rating, and so on. Following the same steps, I scrapped the dataset by first collecting the information of each house cards for the Omaha metro, then I used the links came with each house card to collect more detailed data for each house.

The scraped raw data was quite messy, for example, the heating and cooling features have many different unique values, besides, they are text data, so there are also problems come with the abbreviations, uppercases, lowercases and so on. There are also different units for lot sizes (acres and square foot), and I also witnessed a lot of missing values, for example, more than 80 percent of the data for feature “major remodel year” are missing. Therefore, I need to clean the data. Overall, I performed 4 types of data preprocessing as listed below.

1. I dropped those features which are not necessary or having too much missing values (for example, “major remodel year”).
2. I filled the missing values for some of the numeric features. For example, the missing number of bathrooms and bedrooms were filled with their corresponding median values.
3. I unified the categorical features. For instance, basement variable has 19 different values, “finished”, “full”, “yes”, “unfinished” and so on. I divided these into four different values according to their meaning, “basement (not specified)”, “finished”, “partially finished” and “none”. Which became much clearer while retain the essentials of this feature.
4. I fixed the outliers. Two strategies were used, I dropped some of them, and I replaced some others with the corresponding median value.

For exploratory data analysis, boxplots were created for each variable in order to assess the distribution of the data. Scatterplots were created between each numeric variable and target variable in order to find the correlation between them. Boxplots were also created between each categorical variable and target variable in order to find the correlation between them. To select features, I plotted the pairwise relationship of the features and correlation matrix, I used them to assess the multicollinearity of the features, I dropped the highly correlated features, because correlated features will potentially introduce bias to the models.

For the target variable, the histogram chart was generated to assess the distribution of the data, which did not appear to be so normal. I further confirmed it using the probability plot, which clearly deviated from the linear line representing a perfect normal distribution.

For data modeling, firstly the data was divided into test and training data sets. Since the task of this projects is a typical regression problem, several regression models were trained, such as the multiple linear regression, support vector machine regression and random forest regression. To evaluate the models, I calculated the R-squared value and mean squared error. Then the model with the best performance will be deployed to predict the housing prices.

**Results**

To find out the correlation between independent variable and dependent variable, I used scatter plots and box plots.

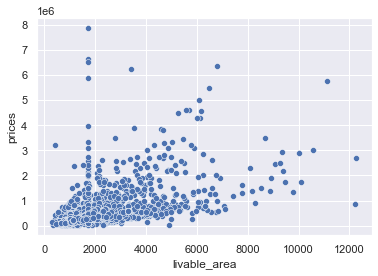


Figure 1: Scatterplot showing the correlation between housing price and livable area.

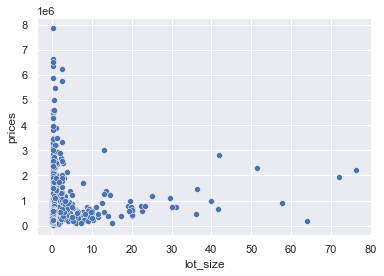


Figure 2: Scatterplot showing the correlation between housing price and lot size.

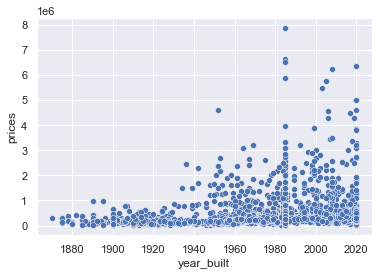


Figure 3: Scatterplot showing the correlation between housing price and year built.

It can be found that livable area had an approximately linear relationship with the housing prices. I think this is in consist with common sense that the more livable area a house has, the higher of the house price will be. And I cannot witness clear correlations between the “housing price” and “lot size”, “housing price” and “year built”.

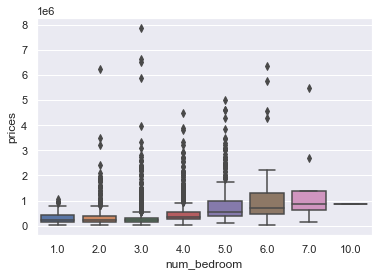


Figure 4: Boxplot showing the correlation between housing price and the number of bedrooms

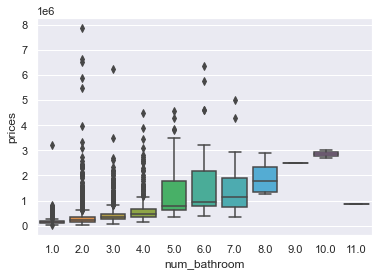


Figure 5: Boxplot showing the correlation between housing price and the number of bathrooms.

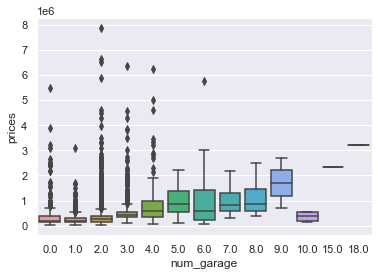


Figure 6: Boxplot showing the correlation between housing price and the number of garages.

It showed that the number of bedrooms, bathrooms and garages can affect the housing prices. In general, the more bedrooms, bathrooms and garages a house has, the higher the price is. This also agrees with common sense, because more bedroom and bathrooms generally indicate a larger living area.

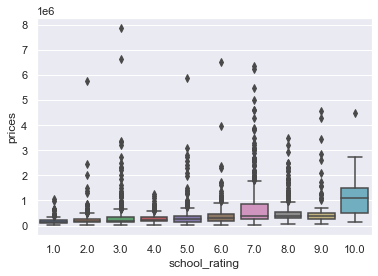


Figure 7: Boxplot showing the correlation between housing prices and school ratings.

The plot clearly shows that the houses in school rating 10 have relatively higher prices.

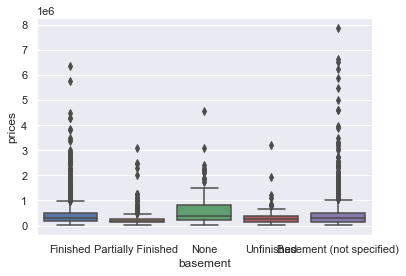


Figure 8: Boxplot showing the correlation between housing prices and basement.

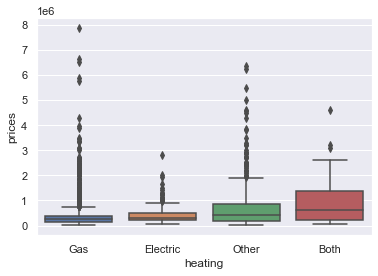
.

Figure 9: Boxplot showing the correlation between housing prices and heating type.

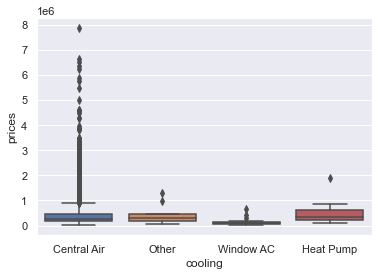


Figure 10: Boxplot showing the correlation between housing prices and cooling type.

It can be seen that “Window AC” has relatively lower housing price.

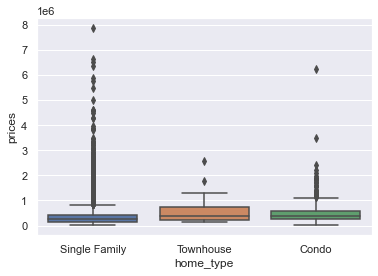


Figure 11: Boxplot showing the correlation between housing prices and home type.

I cannot witness clear correlations between the housing price and basement, heating, and home type.

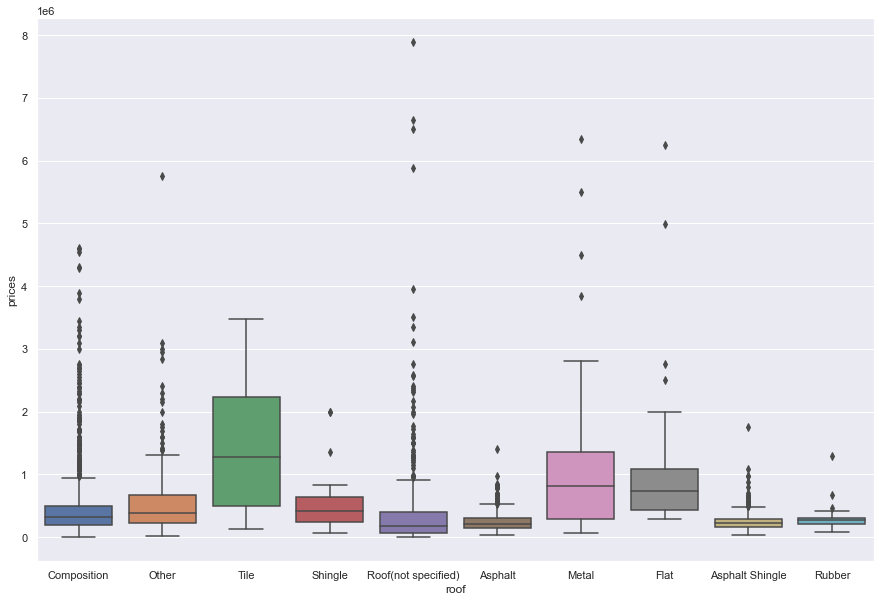


Figure 12: Boxplot showing the correlation between housing prices and roof type.

In these roof types, tile roof, metal roof and flat roof have relatively higher housing prices.



Figure 13: Pairwise relationship of the features.

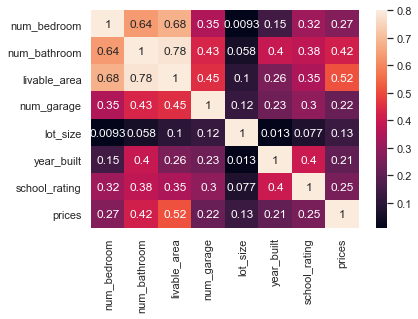


Figure 14: Correlation matrix.

As you can see in the figures above, there are no high correlated features, thus I can keep all these features.

Next, I explored the target variable, the “housing prices”. Histogram chart was created to access the distribution of data. It does not look like so normal. This can be further confirmed from the probability plot which is obviously deviated from the linear line representing a perfect normal distribution.

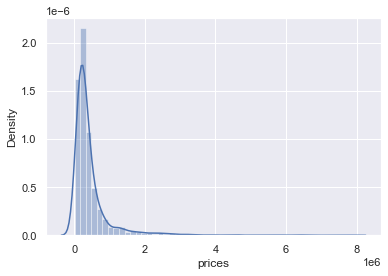


Figure 15: A histogram plot of the housing prices.

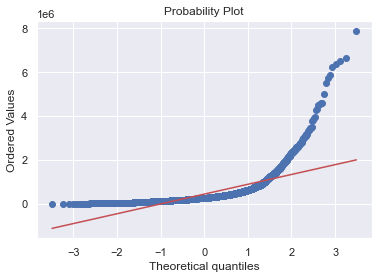


Figure 15: Probability plot of the house price.

After the exploratory data analysis and data preprocessing steps, multiple linear regressors were trained using the train data. The model yielded a multiple R-squared score of 0.38, which is not good. The problem could be the skewness of the target variable, as I mentioned before. So I applied log transformation to the target variable. Then the multiple R-squared increased to 0.55. We can observe something interesting here, the roof flat and roof metal have a large positive coefficient, while the number of garages which presumed to be a correlated feature is not significant, although it has a large positive coefficient. The mean squared error of this model is 0.37.

Graphical user interface, table

Description automatically generated

Figure 16 Feature importance and the corresponding p-values.

Due to factor that the feature “number of garages” is not significant, I removed this feature and trained the linear regression model again. Similar performance was observed, where the roof flat and roof metal held the top two positive significant coefficients, and cooling with Window AC had the smallest negative significant coefficient. The multiple R-squared is 0.54, and the mean squared error of this model ais the same with the last model.

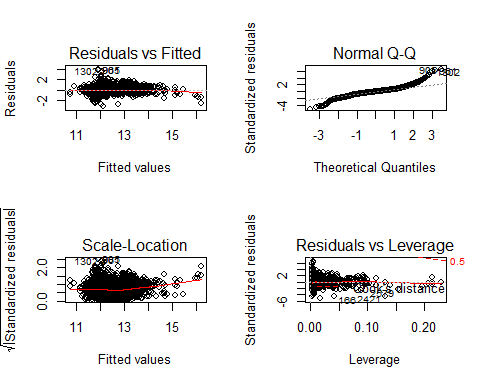


Figure 17: Residual plots of data.

Next, I turned to check the assumption about the residuals. The first graph is a plot of fitted values against residuals. The dots nearly evenly dispersed around zero. Thus, the assumptions of linearity, randomness and homoscedasticity have been met. The second Q-Q plot shows that the dots are distant from the line at the extremes, which indicates a deviation from normality. And all cases are within the dashed Cook's distance line, indicating no clear outliers.

Since linear regression is a fairly simple model, I also tested support vector machine regressor for this project. If directly use the data for support vector machine regression, an extremely large mean squared error will be returned, this is because support vector machine is very sensitive to feature scaling. Thus, I applied log transformation to the target variable again, the mean squared error dramatically decreased to a small value of 0.31. Comparing with the linear regression model, the performance is slightly better.

The last model I used in this project is the random forest regression. I need to apply the same log transformation to the housing price data as I did before. The mean square error obtained is 0.24, much better than that of linear regression and support vector machine.

**Conclusion**

In closing, I scraped a dataset from Zillow.com which contains the house information for about 2000 houses in the Omaha metro area. I performed exploratory data analysis and preprocessed the dataset to trim out some redundant features and fixed the missing values and outliers. The I trained three types of regression models, namely, multiple linear regression, support vector machine regression and random forest regression using R package, the best model performance comes with the random forest model, which has the smallest mean squared error of 0.24 among the three models. Interesting, the model used in this project also predicted a positive correlation between roof types and house price.

For future work, it is noticed the model performance is not quite good at the current stage, and I think there are multiple reasons for this, for example, the model hyper parameters have not been optimized yet, grid search or randomized search with cross validation could be performed in the future to identify the best hyper parameters for the models. Also, more features could be collected to complement the current dataset. It would also be interesting to increase the size of the dataset by collecting house information from east coast and west coast and see the determining factors of housing price vary with the geographical information.

**References:**

1. zillow.com
2. Holden Lewis. Mortgage Interest Rates Forecast. <https://www.nerdwallet.com/article/mortgages/mortgage-interest-rates-forecast>
3. <https://en.wikipedia.org/wiki/Mean_squared_error>
4. Israel Rodrigues. CRISP-DM methodology Leader in data mining and big data. <https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781>