Insights of the Omaha-Lincoln-Des Moines House Prices

Jun Dai

March 28, 2021

1. Introduction

1.1 Background

A home is usually the biggest expense for a common family. A trusted way to monitor the value of the asset is incredibly important in modern life. For decades, people are struggling to get their prediction of a house. It shows hope when machine learning is applied in this area. Zillow is an American online real estate database company dedicated to empowering consumers with data, inspiration, and knowledge. Zillow has data on approximately 110 million homes across the United State. They offer several features including value estimates of homes, value changes of each home, aerial views of homes, and prices of comparable homes in the same area. Also, they provide the basic information of a home, such as square footage, the area of the land and the number of bedrooms and bathroom. This not only gives the users the information needed for their decisions of buying a home, but also offers us large scale of data to do data investigation.

1.2 Problem

Zillow also offers an estimate value of a home known as “Zestimate” based on a range of publicly available information. However, its algorithm is not publicly available and there is always controversy about the accuracy of “Zestimate”. Using data published on the Zillow website, the typical Zestimate error in the United States in July 2016 was $ 14,000. According to Los Angeles Times, Zillow seemingly overemphasizes home square footage as the major metric driving property valuation [2]. Furthermore, Zillow offers national wide real estate information, and it is doubtful that they will design different algorithms for different area. However, situations vary tremendously from east coast to west coast, from north to south. Zillow, themselves, started a competition with a one-million-dollar grand prize challenging the data science community to help push the accuracy of the Zestimate.

Another problem is that ‘location’ has been regards as an important factor for house prices and how people decide which house to buy, either for schools or living convenience, while the original Zillow data does not cover this quite well, how do the neighborhood venues affect the house prices?

1.3 Interest

The interest could from both house buyers and real estate industry. House buyers could know more detailed determining factors for the house price, and real estate industry could learn what to invest in the neighborhood to increase the house prices and make more profits.

2. Data acquisition and cleaning

2.1 Data sources

The Zillow house information was scraped from (https://www.zillow.com/homes/for_sale/) for the cities in the Omaha-Lincoln-Des Moines area. From here I was able to get a dataset with "price", "address" and "link", then use each of the link, I further scraped the addition features for each house, include: 'num_bedroom', 'num_bathroom', 'basement', 'flooring', 'heating', 'cooling', 'livable_area', 'num_parking', 'num_garage', 'lot_size', 'home_type', 'roof', 'year_built', 'year_remodel', 'school_rating', 'latitude', 'longitude'.
Then, use the latitude and longitude data, I complemented the previous data with the nearby venues for each house using Foursquare API, specifically, the number of different venues near each house is added to the dataset.

2.2 Data cleaning

The scraped data is quite messy, it has different unit for area related data, such as 'square feet', 'sq. ft.', 'acre', etc. re are several problems with the datasets. There are "one", "two", "three" etc. for the number of bathrooms, garages, and all these need to be converted to numerical values. There are also a fair number of missing values, as shown in Figure 1. Considering the percentages of missing value in longitude and latitude are low, I simply dropped these rows. However, there are a big percentage for the rest features, so I decided to fill the null value data with median value.

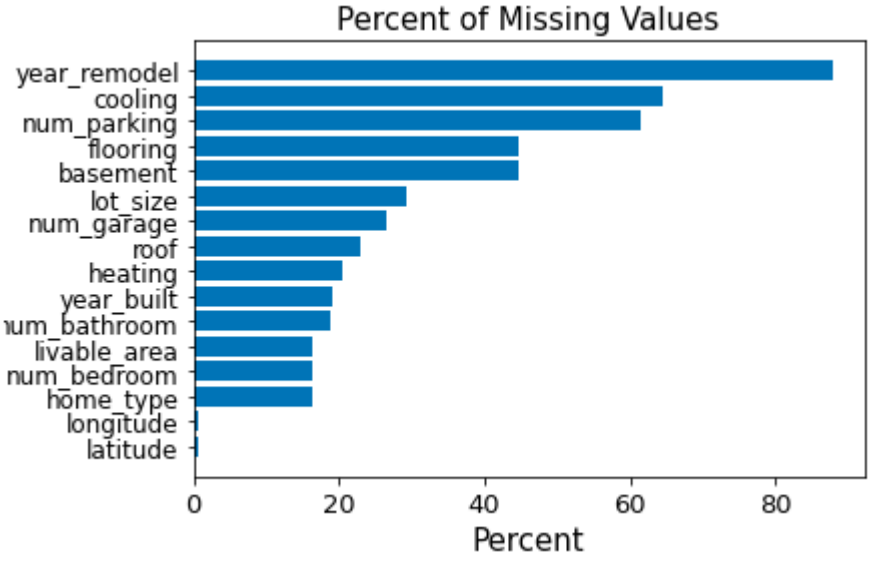

Figure 1. The percent of missing values in the scraped Zillow data.

For School rating, there are several random symbol school rating value, these rows are dropped. For "home_type", as shown in Figure 2, there are a few houses labelled as "apartment", "multiple occupancy", "other", or "mobile", which are also dropped because I want to focus on

single houses, in this sense, only house types of “Single Family”, “Condo” and “Townhouse” are retained.

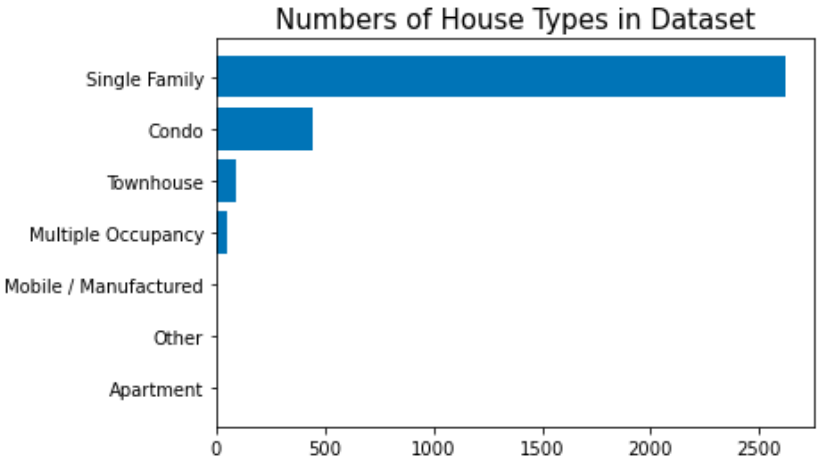

Figure 2. The bar plot of the number of different houses.

For Roof, basement, flooring, address, links, year_built, year_remodel and cooling & heating, there are also some random symbols in these feature and fair amount of null values. I drop the rows where have random symbols and fill the blank features will the corresponding median values. Then I dropped “address” and “links”. Additionally, we also change “flooring” into a dummy form.

For the nearby venues, I used Foursquare and obtained the nearby venues in 3200 meters, which is about 2 miles. Then I count the number of venues and add it to the data set.

The final dataset has the following geographical distribution and has a dimension of 3044*485.

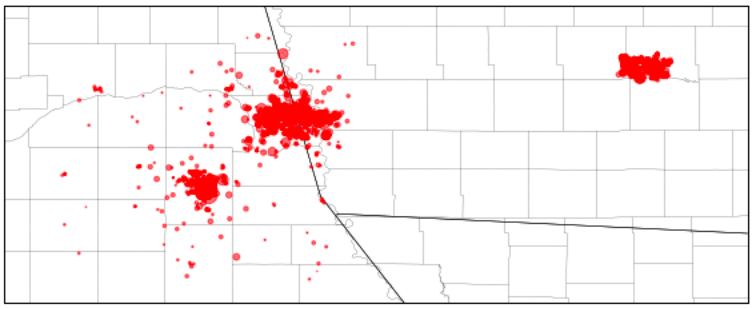

Figure 3. The location of houses used in the dataset.

The dataset is quite sparse because of the number of venues are very geographical related, as shown below, the number of records for each variable which has zero is quite large.

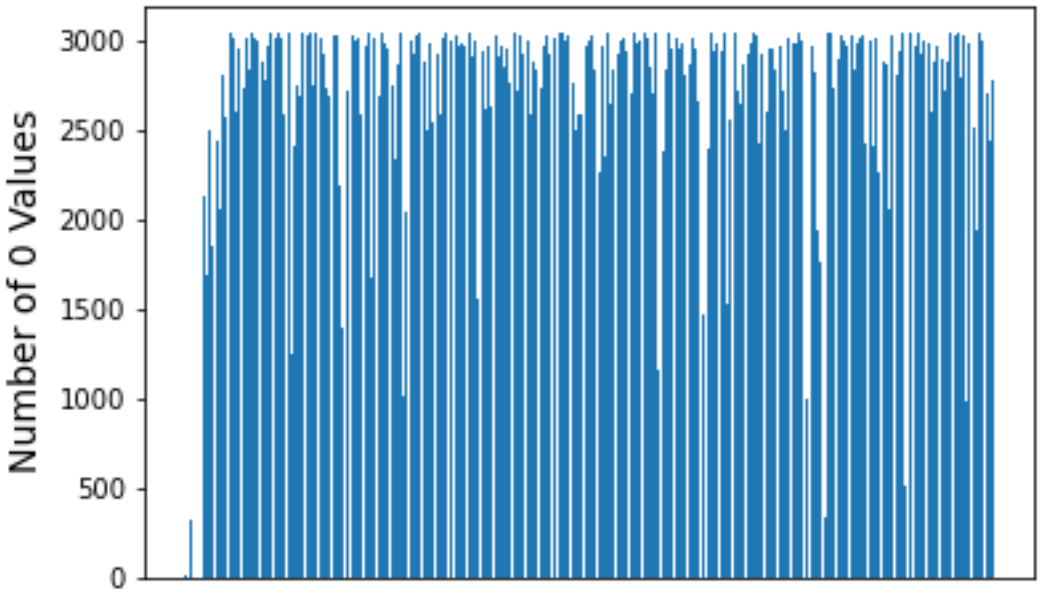

Figure 4. The number of zero data each variable has.

Also, some of the variables are quite correlated, such that the multicollinearity should be considered when choose a model, for example, simple form of linear regression might not a good choice because the data has multicollinearity.

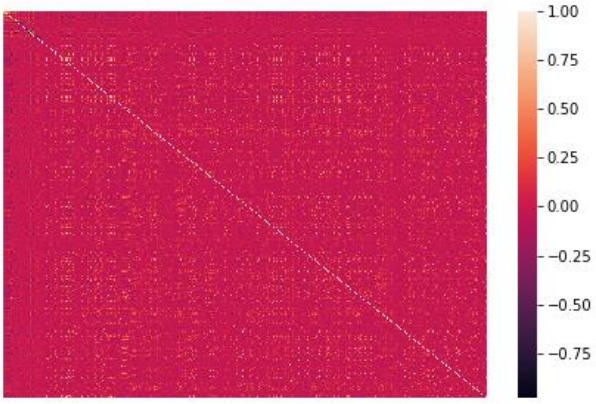

Figure 5. The correlation matrix of all features.

3. Methods

As mentioned before, the features in the dataset are correlated, therefore, the correlation needs to be penalized, in this sense, I choose Lasso and Ridge form of linear regression models to start with, starting with simple models is a general guideline. I split the dataset into training (0.75) and testing (0.25) and performed hyper parameter optimization on the training dataset using grid search cross validation, for Lasso and Ridge, the parameter space is simply the 'alpha'. Model performance is evaluated on the hold out testing dataset. The features used for Lasso and Ridge were normalized before being feed into the machine learning models, because Lasso and Ridge are sensitive to the absolute values, large value could carry a higher weight.

I also used random forest regressor, random forest is not very sensitive to categorical data as well as the range of data, it is also less sensitive to multicollinearity as compared to linear regression. Similar evaluation strategy was used for random forest, except random search cross validation search was used instead of grid search, this is because the size of the hyper parameter sets is much larger that the Lasso and Ridge models, I can not afford the search cost of grid search. Recursive feature elimination technique was also used to automatically select the features, since among the 484 features, some of them may simply make negligible impact.

4. Results and Discussion

Lasso regression performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the regression. Grid search shows an optimal "alpha" of 0.01 with a training r2 score of 65%. While the performance on test set gave a r2 score of 74%. The discrepancy between training and testing performance hints it would be better to use a more complex model. The feature importance of Lasso model is plotted in Figure 6.

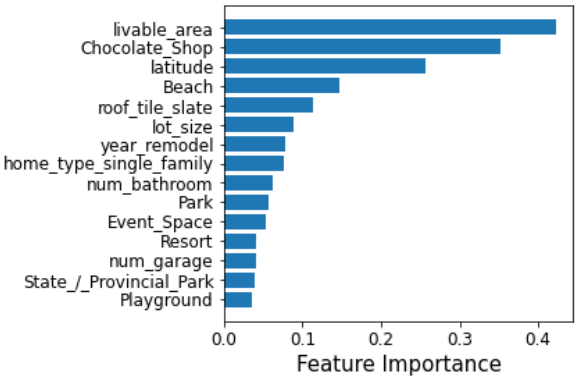

Figure 6. Feature importance plot of the Lasso regression model.

It can be seen livable area appears to be the most important, followed by the number of "Chocolate shop", and "latitude". "livable area" is the most important factor used in the Zillow model, so it is expected that "livable are" is an important feature. "Latitude" and "longitude" represent "location", so it is understandable that "latitude" holds a big factor. However, the number of "Chocolate shop" sits at the second most important is very surprising. Since unlike the other feature "beach", "chocolate shop" does not generally believed to be a sign of luxury.

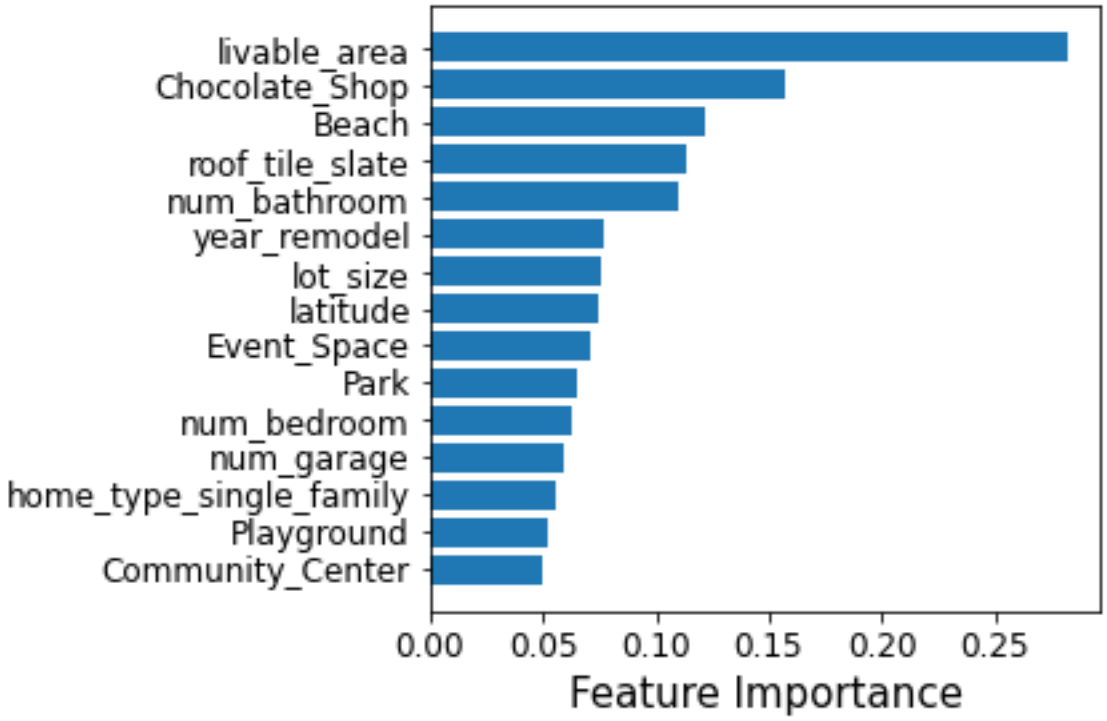

Figure 7. Feature importance plot of the Ridge regression model.

Similar to Lasso, grid search provided an optimal "alpha" of 500. The training r2 score is 65% while the r2 score on test set is 75%, which leads to the move of using a more complex model. For feature importance, it is similar to Lasso in general, "Chocolate shop" and "beach" show up in top 3 while livable area holds the most important feature.

Random forest regressor was used as a more complex model as compared to Lasso and Ridge, it is an ensemble model and not so sensitive to categorical data. However, the hyper parameters have a much larger set to be tuned as compared to Lasso or Ridge models, so I chose random search cross validation instead of grid search cross validation, the champion model has a max_depth of 90, min_samples_leaf of 2, min_samples_split of 5 and n_estimators of 733. The training r2 score is 81% and on test set, it is 89%. Which is more consistent as compared to Lasso and Ridge. The feature importance is plotted below in Figure 8. Where we can see livable area, latitude, longitude at the top 3. It is consistent with common sense, and the reason why longitude is less important than latitude might be because the houses scraped actually have quite close longitude. Interestingly, "Bar", "Chocolate shop" and "book store" are found among the top 10 important features.

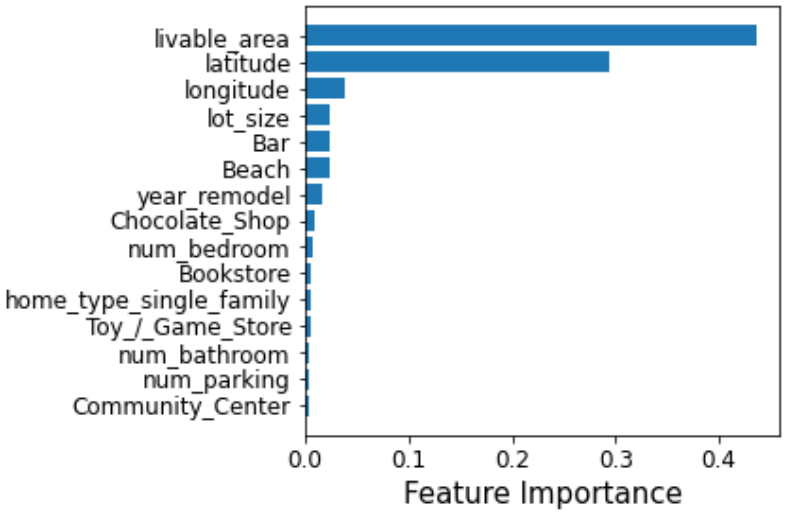

Figure 8. Feature importance plot of the Random Forest regression model.

Although by using random forest, the model performance is much better, however, the performance on training set (81%) and testing set (89%) still differs. This might be a reason of too many features are feed into the model, considering the size of features (484), I chose to use the recursive feature elimination method, and I set the target features to 0.3 of the original (145). Then I did a similar random search cross validation on the training set to find the optimal hyperparameter. Then evaluate on the test set. It appears the training r2 is 81.7% and the testing r2 is 81.0%, which is quite balanced. The feature importance is plotted in figure 9. For those non-zillow features, one can see “beach”, “toy/game store”, “bar” and “chocolate shop”.

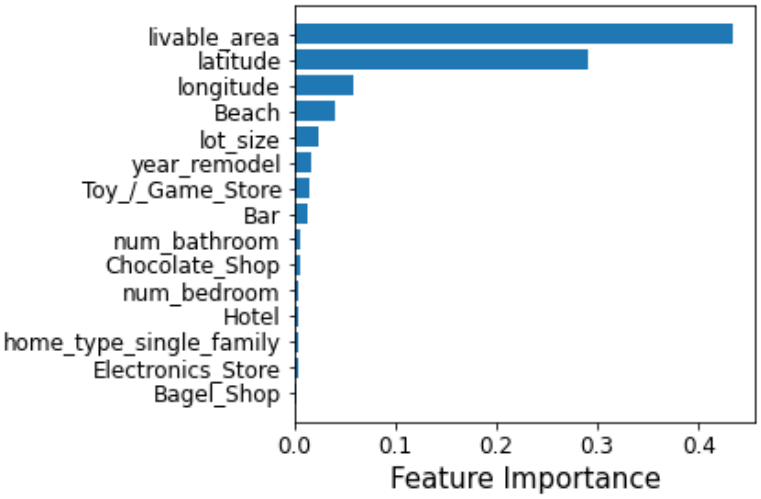

Figure 9. Feature importance plot of the Random Forest regression model with recursive feature elimination.

5. Conclusions and Future directions

According to the result, people in Omaha-Lincoln-Des Moines Area care conventional features such as the liveable area, location, lot-size, year_remodeled (or built) in a descending order in general. Interestingly, “beach” and “chocolate bar” emerge as the two interesting venue related features, although “beach” is understandable, “chocolate bar” is not so common, could be an investing focus if a new neighborhood is under construction. Also, give houses of similar price range, house buyer could take a look at the nearby venues such as “chocolate bar”, “bar” etc. for an extra estimation of the investment.

For future direction, a deep look at the nearby venues is needed to confirm the finding, because of the COVID-19 pandemic, a lot of small businesses are not function normally, so there might be a chance to have abnormal venues data, even though the venue data is collected from recent Foursquare call. I would also use xgboost and other feature selection techniques to further improve the model performance.