Real Estate in the Social Media Age: A study on Beijing Housing Prices 2009-2011 on Lianjia

Leslie An, Sara Bolf, Camille Little

Rice University, Houston, Texas

December 15, 2019

Abstract

While the traditional metrics such as number of bedrooms do matter, in the modern age there are other concerns for house buyers that are not expressed by those numbers. With that in mind, this report gathered data from Lianjia.com, a real estate trading platform with a social media style presentation operating in large cities in China, and attempted to discover what are the major contributors to a house's final trading price. Linear regression and variable selection techniques were used to relate price and other metrics. Another point of interest in this report is Beijing's social stratosphere in relation to the housing price. As in any city, the districts in Beijing are not equal and the housing prices should reflect that. Techniques such as QDA and random forests were used for this classification. The regression methods used were able to predict housing price with a mean squared error of 0.39, and the best classification technique classified districts with a total error equal to 0.35.

1 Introduction

The housing market has been a point of interest for many statistical studies, but many have stuck with the most observable metrics when evaluating housing prices in a certain urban area: how big the house is, how old it is, or how many bedrooms it has. However, the physical attributes do not tell the entire story. On one hand, there are factors not expressed by the house itself, such as the neighborhood or its proximity to one's work location. On the other, some consider buying a house an investment itself, and thus the social media aspect of a website such as Lianjia has implications not dissimilar to a stock market. All traders can see the trades that have transpired and the current bid/ask price of the property of interest, as well as how many people are currently following this potential transaction. This platform turns the previous obscure business of real estate into a rather transparent system, and the game of trading thus changes as a result. Simple "new" metrics such as district or followers allow investors to peer into the mind of a potential buyer when looking at houses, going beyond the basic attributes.

The housing data in Beijing provides the chance to study such a system of real estate trade, and the possible impact it has on an established market. Currently, there is no such platform in the United States, but in the age of social media, it is conceivable that a similar service will be implemented in the near future. In addition, since Beijing is one of the largest urban areas in the world, this data provides general insight about the housing market in a metropolis. House buyers and policy makers alike would find such a study insightful, as it focuses on the inner workings of a fine-tuned industry, one that is an integral part of society. Sociologists interested in urban inequality and the social structure of Beijing would also find this study interesting, as it examines the different districts of Beijing and their possible relationship with the housing market.

This problem has been attempted by many statisticians before, although similar to most economic behavior puzzles, the housing market can never be truly "solved." The main mystery that makes social science problems difficult is the lack of any controlled experiment: unlike the disciplines in natural sciences, social scientists can only observe what is given, and infer from an imperfect environment what the results mean.

Reference literature in this area is too numerous to list in the report, but many have studied the relationship between certain characteristics of a given property and its final trade value. However, this

report is unique in the sense that no other literature has considered the specific social media impact on housing prices. Furthermore, this report will attempt to evaluate the level of equality (or inequality) between the districts in Beijing based on the activities in its housing market, which again has not been referenced before.

The report focuses on two key questions:

1. What are the effects of each housing metric on price?

2. Can data accurately classify the houses into the correct district based on their prices and other housing metrics?

The report attempts to answer these questions by using the Beijing housing data collected on the website Lianjia. The raw data was compiled by a third party and uploaded on Kaggle, but all data was available on Lianjia itself, which catalogs all of its trades and timestamps its data for each particular post/transaction.

The first question can be answered with numerous regression techniques, since the report is concerned with predicting the correct price of a house based on data available to the public. The second question is more nuanced, and multiple approaches could have been taken. Ultimately, the report focuses on classification since any other approach would introduce human bias and therefore inaccuracy.

To infer prices via regression, the report starts with a simple ordinary linear regression. The report then employs ridge and LASSO regression in order to gain a more comprehensive view of the relationship between price and the other housing variables. Additionally, employing more than one regression method allows for the use of different techniques for the selection of the relevant variables.

In order to classify districts based on other housing variables, the report employs four classification methods. The report uses classification trees, since trees are easy to interpret and provide insight on how the data can be segmented. Since the data set is large, the report prunes the tree for easier interpretation. Random forest is implemented for a more robust approach, as random forest tend to outperform the more situational bagging. Multinomial logistic regression and quadratic discriminant analysis (QDA) are also used, mainly for comparison purposes.

After the above analyses were performed, it was found that the regression was mostly successful. All methods, including the OLS, returned a mean squared error of around 0.39 on the standardized data, meaning that the regression can predict the price of a house given the metrics with decent accuracy. Out of the methods employed, elastic net seems to have performed the best, having the lowest MSE. The analyses also resulted in an interesting discovery about the LASSO regression. Given the selected penalty terms for lasso, there were no variables omitted; all variables were considered important and included in the regression, barring the variable omitted due to faulty data collection or bad data quality. Overall, the main driver for the price seems to be renovation condition.

With regards to classification, random forest performed the best out of the 4 methods used, having the lowest total error. However the error rate is still relatively high at 35%. In other words, the classification methods used in this report cannot reliably pinpoint a location of a property given its attributes. This is somewhat expected: in a city that is more homogeneous, there should not be significant differences in the quality of the housing as to indicate which district a particular house belongs to. This corresponds to our linear regression results, in which most districts were not deemed to be significant.

The report begins by showing the exploratory steps taken when the data set was first acquired. In this section, some of the preliminary plots made for simple interpretation are included, as well as some of the prototypes of techniques attempted in the early stages of the project. Visualizations are also included to help the readers understand both the data and the methodologies. The report then walks through the data analysis in detail, including all methodologies and all relevant plots and tables thereof. For most techniques used, tuning parameter selection is required, as well as cross validation if necessary. The results from the analyses are provided, as well as any plots and tables produced if they are considered interpretable. The strength of the results is measured using the test data set, and presented in the form of tables. Lastly, a conclusion is made concerning the questions presented above, based on the exploratory data, regression, and classification analyses. Additionally, an appendix including all figures too burdensome or unaesthetic to include in the main report is provided.

2 Exploratory Data Analysis

In order to answer the key questions posed in the introduction, two analyses were implemented. For the first question, the focus of the analysis was the impact of certain housing metrics on the price. Therefore, the method chosen for this analysis was a regression on price. For the second question, the analysis centered around the classification of districts based on various housing metrics. To perform

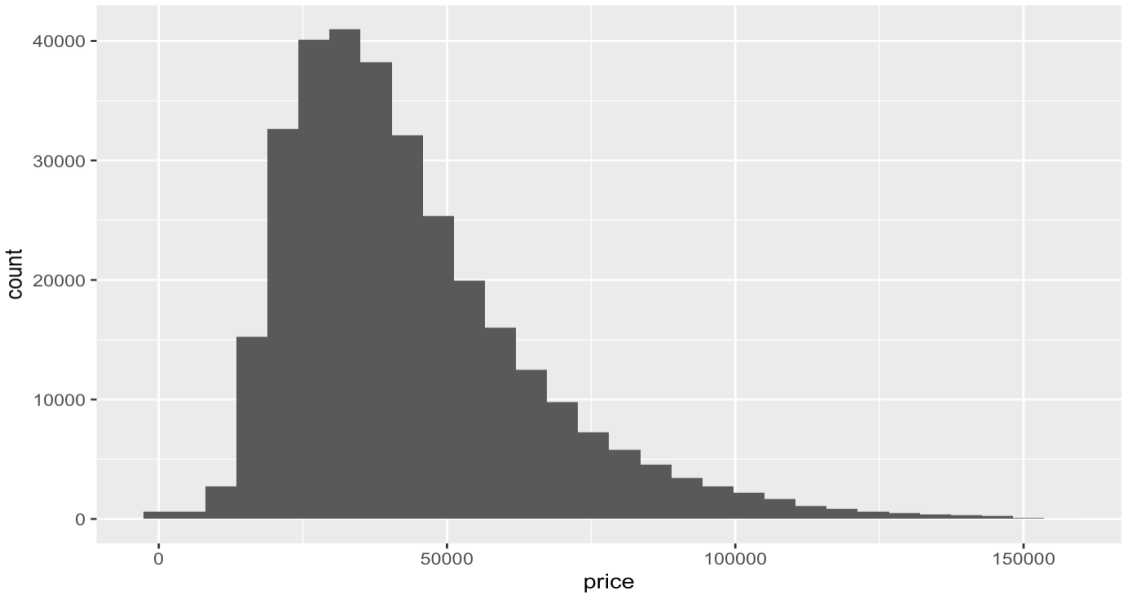

Figure 1: The histogram for all housing prices per square meter in our data. Perhaps as expected, the housing price skews to the right with an average of 43495 Rmb and a median of 38726 Rmb. The most expensive price was 156250 Rmb per square meter. 75% of data falls under 53000, and above 28000.

this analysis, four methods of classification were implemented: multinomial logistic regression, quadratic discriminant analysis (QDA), decision trees, and random forest. Then, the test total error using DATA2 was computed for each method in order to compare the accuracy of the methods.

There are several variables that were removed at the beginning since the variables are either not relevant or they are badly documented.

Trade ID was removed since it is irrelevant. Days on the market is removed due to the fact that most entries on the variable is either 1 or NA, which provides no insight and only makes operations difficult. Total price was removed due to it being a duplicate of price, and including both will not be conducive to correct analysis.

As with all other statistical analysis, the first order of business is to plot out the variable of interest and see if there are any anomalies detectable to the human eye. Figure 1 is shown above.

The skewedness is more or less expected from a price data set since the data has a lower bound but no higher bound. While there are some outliers, they were very few and were included.

Since a histogram for categorical data is unnecessary and not helpful, the next operation done was to reduce dimensions and try to see if there are noticeable patterns there. A principal component analysis is done on the continuous data, and the cumulative variance explained is plotted below. Figure 2 is attached in the next page.

In addition to the plots shown below, other plots were made during the exploratory data analysis phase of the project. The price by days on market (DOM) was plotted in attempt to determine a relationship. When plotted, there was significant fluctuation in the DOM variable, regardless of the price. Additionally, DOM has substantial NA's, so the variable was extremely difficult to model. As a result, this relationship was not further explored, and the DOM variable was removed from the data set.

In another plot, the distribution of construction year for each building type over a seven year period was plotted. Most of the building types had a relatively even distribution of construction times except for building type 2. While this relationship could have been interesting to explore when predicting price or district, the building type variable does not have a description on Kaggle making the variable uninterpretable. This relationship was not further explored.

After the cursory PCA, data exploration took a somewhat divergent route between the two different questions that needed answers. For the price model, more analysis on each individual variable was done to make sure no influential points were outliers. For the classification of districts, some preliminary classification was done to see the effectiveness of the techniques. It is here that the report discovered that the longitude and latitude included in the data would perfectly classify each district. This discovery also allowed us to find each district's real names, since they were not included in the data set. After the classification is done, it is possible to return to the real names and consult with conventional wisdom

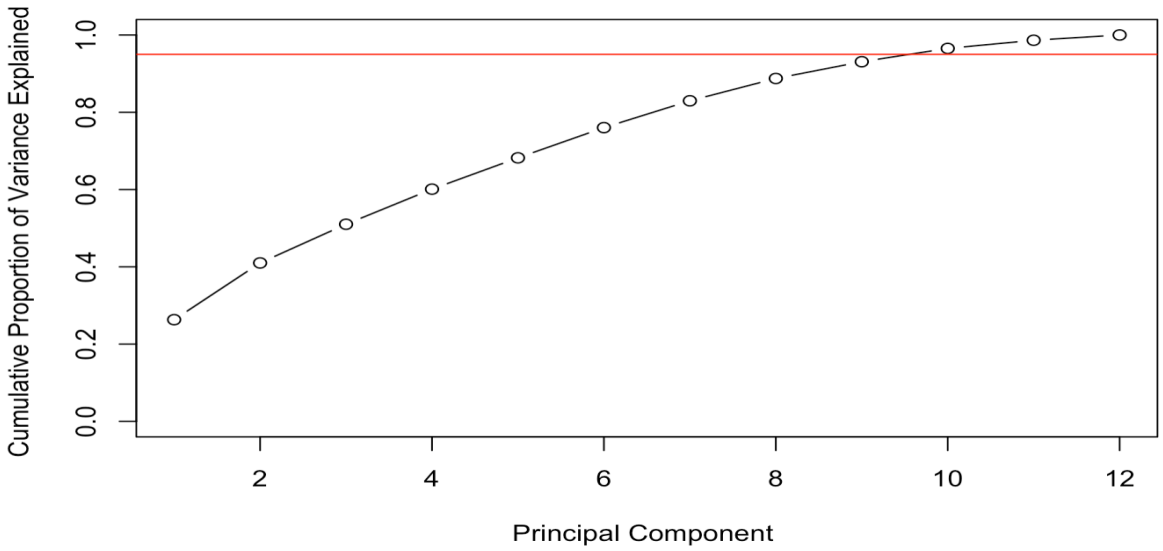

Figure 2: The cumulative variance explained against the number of components used. There is no clear shoulder, which is less ideal, and to explain 95 % of the variance, 9 components needs to be used. While 9 is less than the variables in the data set, it is still a lot and the relatively even spread of variance is somewhat concerning.

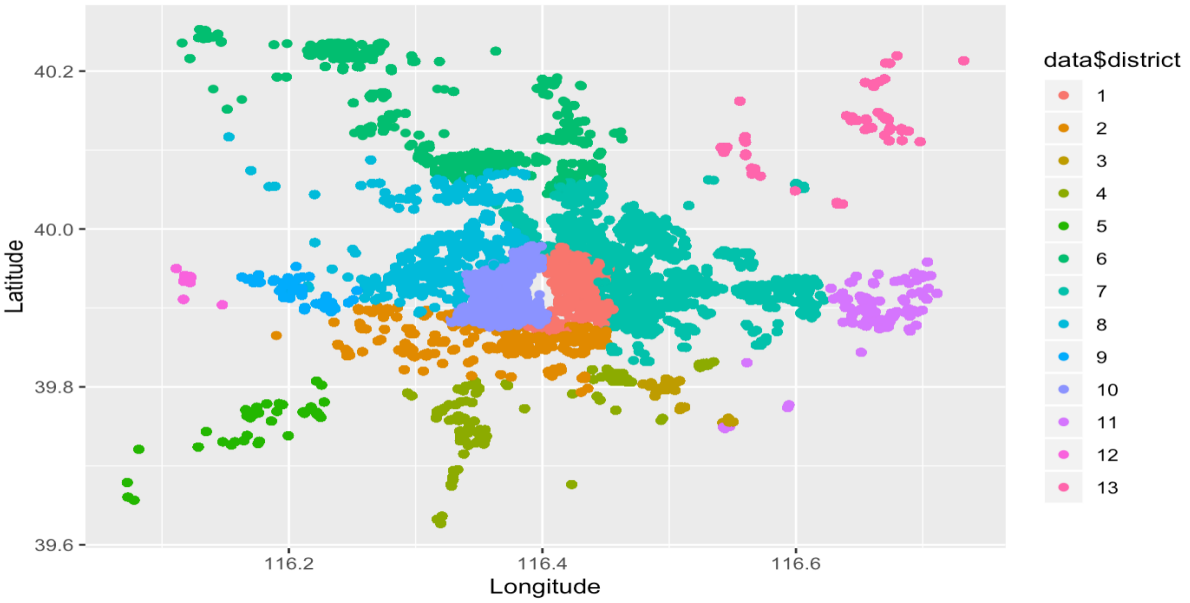

Figure 3: The districts of Beijing plotted against longitutde and latitude. The initial oversight resulted in extremely accurate classification since the physical location of each district is bounded and fully explained by these coordinates. However, the effort was not fully wasted since the data itself did not name the districts, merely gave them ambiguous numbers. With this reconstructed map of Beijing, it is possible to now look up the name of the districts.

to see if the discoveries made here agree with the real world. A plot of the outlines of the properties in Beijing is attached in Figure 3.

3 Analysis

In order to answer the key questions posed in the introduction, two analyses were implemented. For the first question, the focus of the analysis was the impact of certain housing metrics on the price. Therefore, the method chosen for this analysis was a regression on price. For the second question, the analysis centered around the classification of districts based on various housing metrics. To perform this analysis, four methods of classification were implemented: multinomial logistic regression, quadratic discriminant analysis (QDA), decision trees, and random forest.

3.1 Regression

Before performing regression on the variable price, several variables were removed. First, the variables cid (Trade ID) and trade time were removed because they were not relevant in the analysis. Next, days on market was removed because the majority of the values were missing. Building type was removed because Kaggle does not define the various building types, making the variable uninterpretable. Building structure was removed because the most of the samples were categorized into the cement building structures, so it did not make sense to include this variable in the regression. Construction time was removed because in China, new buildings are not constructed very often, so construction time often marks the start of a renovation. As a result, the variable is not a good representation of when a building was constructed. Lastly, longitude and latitude coordinates were removed because they are similar to district and it is not necessary to include all three variables in the regression.

In total, four regression techniques were performed. First OLS was performed. Followers, renovation condition, and community average income were amongst the variables designated as statistically significant. Lasso, ridge, and elastic net were then performed. The table below shows the results from all four regressions. Lasso, elastic Net, OLS, and ridge all have similar mean squared errors.

	Mean Squared Error	Lambda
OLS	0.39565	NA
Ridge	0.39573	0.004866
Lasso	0.39555	0.000165
Elastic Net	0.39552	0.000316

Table 1: Table of mean squared errors and lambdas for OLS, ridge regression, lasso regression, and elastic net. The lambdas were calculated using DATA1 and the mean squared errors were calculated by performing the four regression models on DATA2. AS seen in the table, elastic net has the smallest mean squared error out of the four regression models, however the errors for all of them are very similar. All of output lambdas are small.

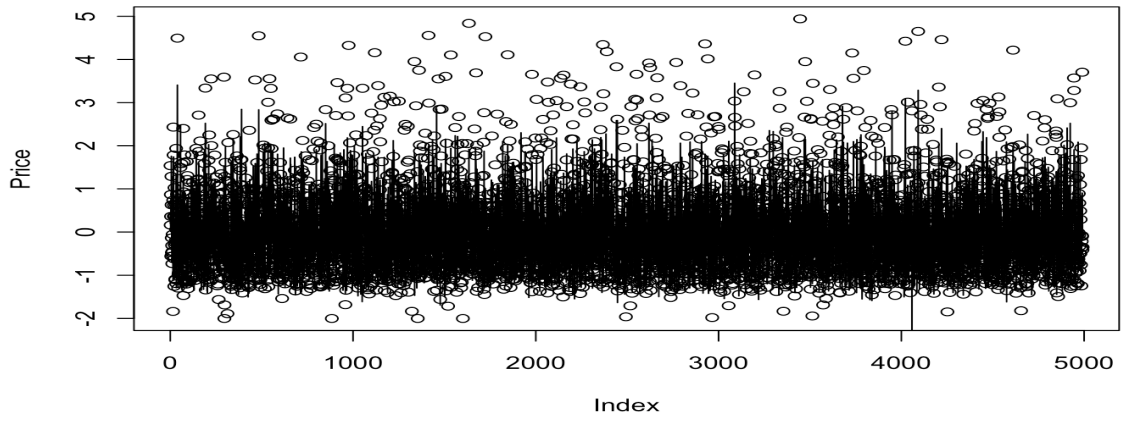

Figure 4: The above plot shows the goodness of fit from running the elastic net model on DATA1. While the plot is very difficult to interpret, it is hardly surprising; there were a lot of data points, and a lot of them cluster around 1 standard deviation away from the mean.

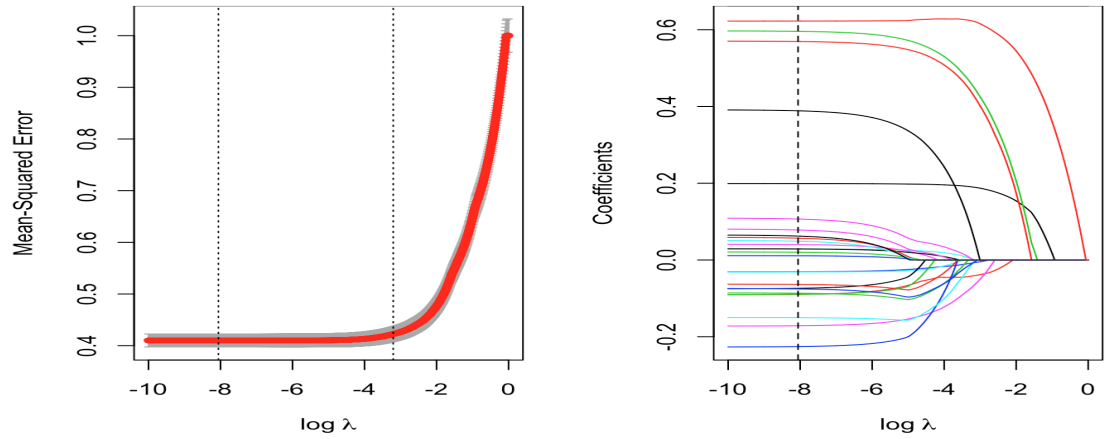

Figure 5: The above plot shows elastic net results from the running the elastic net model on DATA1. The plot on the left displays the total error in response to lambda. As can be seen in the plot, the very small lambdas minimize the mean-squared error. This small mean squared error is reflected in table 1. The plot on the right displays the beta of the variable. The beta is being pushed to zero as the penalty term increases.

3.2 Classification

Before performing classification on the variable District, it was necessary to remove certain variables from the data set. The variables that were chosen to be removed were the same as some of those removed in the regression analysis. Again, trade ID and trade time were removed because they were irrelevant to the classification analysis, and the days on market variable was removed because of the prevalence of missing values. Latitude and longitude were also removed from the data set. During EDA, it became obvious that latitude and longitude had a strong relationship with district, as houses with coordinates close to each other are probably in the same district. Therefore, it seemed best to remove these two variables, so that the analysis could focus more on variables whose relationship with district were not as obvious. Construction time was the final variable removed, for the same reason given in the regression analysis portion.

Since the original data set was so large, all of the analyses were performed on a subset of the training data set, DATA1. The first classification method implemented was multinomial logistic regression. The second method was QDA, but some changes had to be made to the data set before implementing this method. Specifically, the categorical predictor variables had to be removed, as well as the kitchen variable. The third method of analysis was classification trees. The pruning parameter was chosen with 10 fold

cross validation. The final method performed was random forest, with m, the number of candidates for each split, chosen to be the square root of the number of predictors, or 4.

The results of the classification tree analysis can be seen in Figure 6. The average community income is the most important variable in the classification of districts. Specifically, lower average community income values, considered as those less than $\$60,000$, are associated with districts 6 and 7, while higher average community incomes are associated with districts 7,8, and 10.

Table 2 shows the results of the random forest analysis. Based on these results, the average community income is found to be the most important variable for district classification. The square footage of the house is also an important variable. Conversely, the number of followers is less important based on the mean decrease in accuracy, and the number of kitchens in a house is less significant based on the mean decrease in node impurity.

A randomly chosen subset of DATA2 was used as the test data set, and the accuracy of the classification methods were measured using the test total error. The most accurate method was random forest, with a test total error equal to 0.354. Multinomial, QDA, and trees were generally not that accurate, with around 50% of their predictions being wrong.

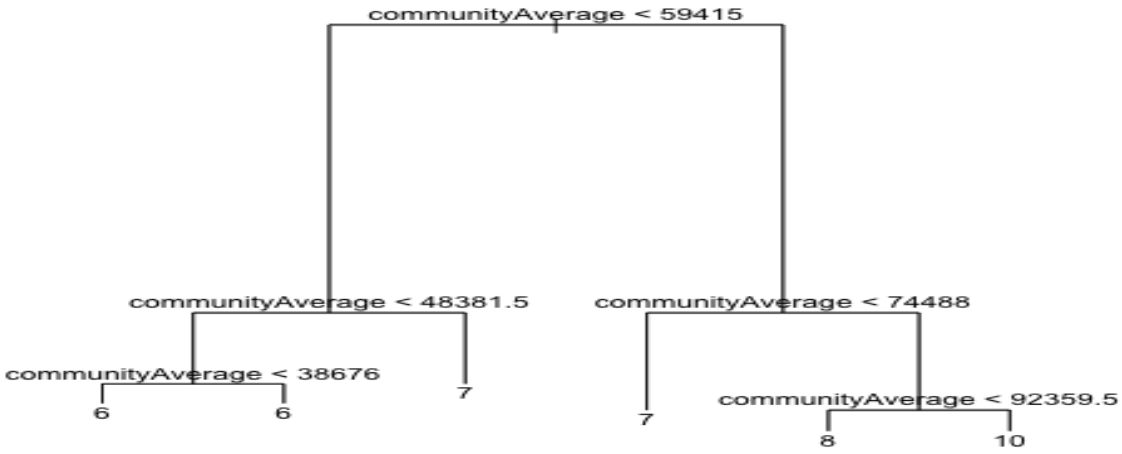

Figure 6: The pruned classification tree for district, with the number of terminal nodes chosen by 10 fold CV. All the splits in the tree are done on the variable community average, which implies that this variable is the most significant in the classification of district. Specifically, the splits in the tree are dependent on the size of community average. The terminal nodes 6 and 7 are associated with smaller values of community average, with the nodes 7,8, and 10 are associated with larger values of community average.

	MeanDecreaseAccuracy	MeanDecreaseGini
followers	43.107	542.904
price	78.410	915.681
square	138.359	925.318
livingRoom	67.258	213.678
drawingRoom	50.316	145.450
kitchen	56.298	30.038
bathRoom	44.582	92.151
buildingType	80.504	193.901
renovationCondition	48.511	244.846
buildingStructure	90.576	162.357
ladderRatio	90.018	473.200
elevator	46.395	89.658
fiveYearsProperty	49.622	135.893
subway	83.090	144.640
communityAverage	234.902	1994.785

Table 2: Table of Variable Importance for DATA1. The first column provides the mean decrease of accuracy in predictions if that variable is excluded from the model. The second column gives the mean decrease in node impurity for a split over that variable. Based on the table, the variable community average has the largest mean decrease of accuracy, while followers has the smallest. The mean decrease in node impurity is again largest for community average, and smallest for kitchen.

	Test Total Error
Multinomial Logistic	0.513
QDA	0.544
Tree	0.534
Random Forest (m=4)	0.354

Table 3: Table of Test Total Error using DATA2 for multinomial logistic regression,QDA,decision tree, and random forest. The classification accuracy for multinomial logistic regression, QDA, and decision tree are similar, all having a test total error around 50%. Random forest with m=4 is the best method based on classification accuracy, as its test total error is much lower than the other three methods.

4 Conclusion

4.1 Regression

The test mean squared error of all the regressions were very close; they all hover just above 0.39. However, the elastic net with 0.7 α provided the lowest MSE. All three elastic net methods selected λ values that were relatively small, and the lasso method did not choose to exclude any variables. This indicates that the variables that made it past the initial pruning process from being statistically faulty or bad in quality were deemed significant enough.

The regression beta focus most on the renovation status. It would appear that the condition that the house comes in is the most telling of its final price per squared meter. This makes intuitive sense, since renovation perhaps best indicates the amount of per unit value for a house buyer. Interestingly enough, the total size of the house is negatively correlated with its squared meter price. This is likely due to the fact that a larger house would sell for more in total, and real estate agents sometimes would lower the price per unit in order to entice potential buyers. As a result, the number of rooms in a house seem to have little effect on the price, as well as varying in terms of positive and negative values. The follower aspect of the data set seem to have a positive impact on the final price, but the effect is relatively small. When the information is widely available, it is difficult to increase the value of a property by a lot.

One interesting aspect of the regressions is that the beta terms for the districts were mostly deemed non-important. This makes intuitive sense since as long as the house is adequate, its location should not matter. However, there are other conditions that might challenge this notion. Different districts are not exactly the same in terms of their neighborhoods, school districts, and public transportation. However, from the regression used, there does not seem to be a clearly defined, significant relationship.

The fitted line of the elastic net regression is difficult to interpret. Since there were many data points and predictors, the fitted line ended up oscillating a lot over the frequencies. This makes the plot difficult to interpret and it was omitted in the analysis part of the report. The oscillations seem to correspond to the general trend of the data, however. Additionally, with the penalty to the over-fitting, this seems to be the optimal fit, and its errors were smaller than the OLS.

4.2 Classification

On the side of the classification, the four methods chosen had more varied errors, but it would appear that random forest with $m = 4$ is the best. The other methods suffer from some issues, though, so that might explain their weaker performances. For example, since the data set included a lot of categorical variables, which had to removed in order to run QDA,the QDA method was not able to compute predictions accurately.

Some insights did arise from the unsupervised learning methods. For example, in the tree method, the earning of the community around the property was deemed the most important factor in determining where the property lies in terms of districts. This potentially makes sense since there could be large and small houses in every district, but it is common for cities to have varying average income in different neighborhoods and districts. The relatively lower accuracy of the tree method suggest that as a whole, Beijing is not very gentrified or segregated by earning, thus making it difficult to deduce the property's location given the earnings of its neighborhood.

The random forest method yielded a error rate of 0.35, which is much better than the rest of the methods. However, even this technique cannot guarantee a completely accurate prediction of a property's district. However, it is important to note that, due to the fact that there are 13 districts, an error rate of 0.35 is actually quite impressive. For example, the error rate of randomly guessing would be around 0.92.

Summarizing over both the results from regression and classification, it can be said that the districts in Beijing are not entirely heterogeneous in terms of the houses that are on sale in each of them. However, some patterns can be drawn, and the districts are definitely not all equal. This can be expected just from a city's natural progression, however, since the business district cannot be the same as the residential districts. Compounding to that fact is that some districts saw relatively fewer traffic in terms of business, and contributed very little information to the study. The unused hierarchical modeling in fact only included 10 clusters.

In conclusion, this report has found a regression method that predicts the final housing price with decent accuracy, and shows that the per square meter price is best explained by the condition in which the house is renovated. In trying to understand the social structure of Beijing, it can be seen that although there are differences in income and typical housing in each district, there are enough variety so that it is not totally segregated.

5 Appendix

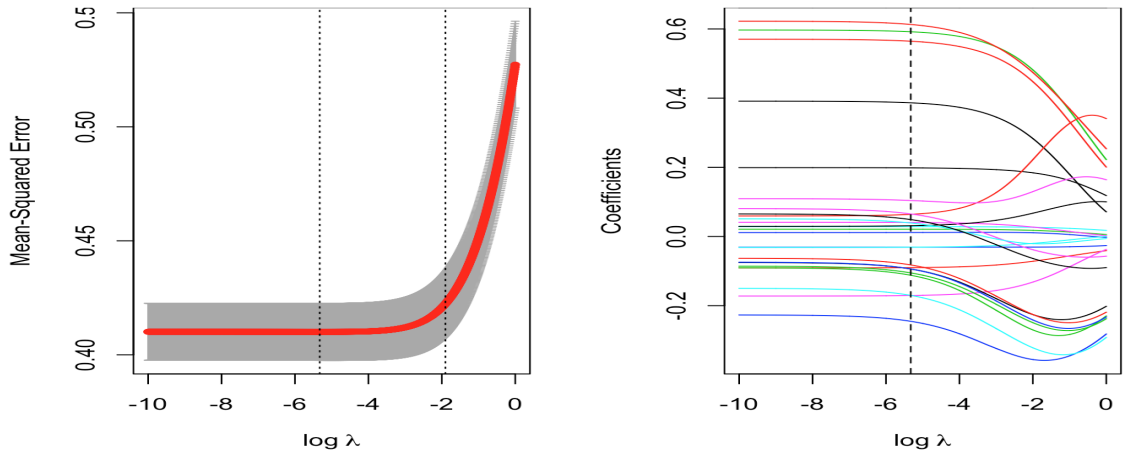

Figure 7: The above plot shows the mean squared error in response to lambda and the beta. The plots were produced by running ridge regression on DATA1.

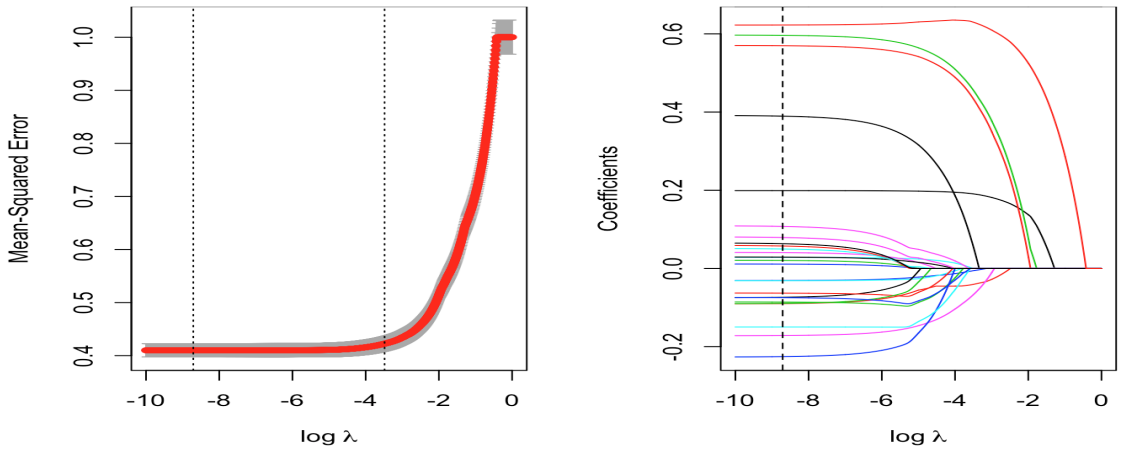

Figure 8: The above plot shows the mean squared error in response to lambda and the beta. The plots were produced by running lasso regression on DATA1.

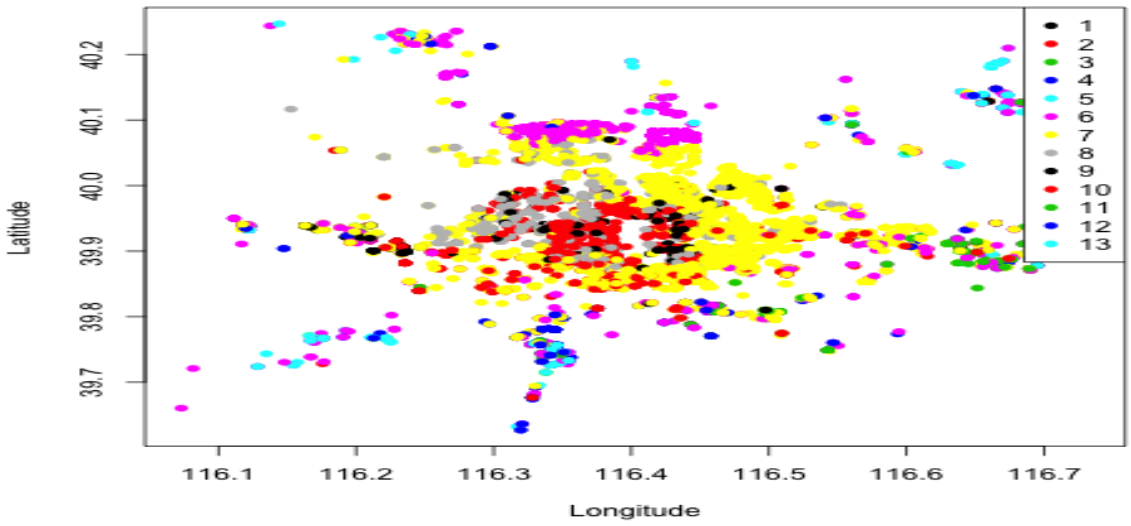

Figure 9: The above plot shows the longitude and latitude variables from DATA2 plotted against each other, colored by the district predicted by the random forest classification.