Analysis of Housing Price in Beijing

Hu Jukai[1]

[1]*WBooth School of Engineering Practice and Technology, McMaster University*, 1280 *Main St W, Hamilton, ON L8S 4L8, Canada*

Abstract

This study examines the trend of Beijing's second-hand housing prices before the global pandemic. It would generate informative predictions for Beijing's second-hand housing prices after the pandemic period, where efforts aimed at economic rehabilitation are ongoing, and the property market (including second-hand housing) is gradually returning on track. This study applied two PCA methods and one machine learning model to test the collected second-hand housing dataset, trying to find potential regression patterns that could be easily interpreted (I used Multilinear Regression, Partial Least-Square, and Random Forest (denote as MLR, PLS, and RF for short) with training, and applied ten-fold cross-validation with repetition to derive R Square, Mean Square Error, and Root Mean Square Error (denote as R2, MSE, and RMSE respectively) respectively. Meanwhile, the trained model was tested with the testing dataset.). With a precise prediction and analysis of second-hand housing prices, not only can we provide a business investment with scientific trending of future second-hand housing prices, but also we can take advantage of the prediction model once we come to purchase houses in the market. This study shows that regression methods with decent interpretability achieve a decent performance on Beijing’s second-hand house prediction, even with the natural tradeoff between the interpretability of a decision algorithm and its accuracy in application. Meanwhile, this study helps to unveil how various housing features are correlated to the total price.

1 Introduction

As the capital of China, Beijing has distinct housing and second-hand housing price trending, and it keeps receiving a lot of public attention since real estate investment is one of the most popular investment modalities with a high return in China. Meanwhile, the economy of every country is strongly associated with real estate pricing, and housing pricing is an indispensable contribution to the GDP (Roughly a 10% drop in property prices leads to a 1% drop in GDP growth in normal times. Because income and future expectations of property prices can cause "serial correlation" in the data (where prices and GDP growth depend on its former self more than supply and demand), a 10% drop in prices can reverberate into a 7–8% decrease in GDP output at the extreme (when prices and GDP radically change course). [1]). Above all suggests that real estate has already been a most research-intensive component in each country's economic system that deserves to extract valuable information.

Besides the necessity of analyzing the Beijing second-housing price, there are other motivations behind this project. One of them is that by combining with the post-pandemic Beijing second-housing price trend in the future, this project will provide us with how correlated the pre and post-pandemic Beijing second-housing market is. During the pandemic period, both the social economy and livelihoods were badly sabotaged in China, resulting in a periodical abnormal real estate economy (The global epidemic heavily impacted the housing market in China. The main reason is that many real estate enterprises stopped selling houses, and local governments implemented home quarantine measures, which affect normal housing transactions (which would affect the second-hand housing market directly). [2] On behalf of the customers' side, the COVID-19 pandemic did not greatly negatively impact their demand and confidence in buying houses compared with the impact of the pandemic posted to house sellers. But still, this situation

is not general enough.). Currently, the Chinese government has loosened control over all sections of society, trying to take society back to the normal track as soon as possible. The economy and customer consuming power is flourishing again, and real estate pricing is returning to fit into the normal trend as before the pandemic began. The paper will make an analysis based on the features of Beijing's second-hand housing provided by owners (the house information will then be posted on lianjia.com, which is a second-hand housing transaction platform). Then, we can decide which feature(s) of the house would be more predominant in deciding the total price and what is the potential correlation metrics between each feature and the total price, which is another motivation. I will try to generate interpretable linear trending to fit Beijing's second-hand housing price from 2009 to 2018 to see if the linear model is explanatory enough to be applied to the post-pandemic second-housing market.

Another point worth mentioning is that the debate over whether China should establish a personal housing property tax has been going on for many years, but the tax has not been introduced. [3] If the property tax is introduced, then this already analyzed pre-pandemic dataset we derived in this project could shed light on analyzing the correlation between newly introduced property tax and other housing features, helping the future second-housing prediction model to be constructed more efficiently.

2 Databased Description

The dataset used in this study is “Housing prices in Beijing, Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com” [4], which is available in Kaggle. Most of the data (Beijing second-hand house information) were collected between 2011 and 2017; It also included some

data on 2009 and 2018, which is proximately the period before the global pandemic kicked off (late 2019). This dataset includes 26 columns and 318,851 rows. Each row represents an observation (house to be sold), and each column is a specific feature of that house. Feature columns include URL, ID, Lng, Lat, CommunityID, TradeTime, DOM (days on the market), Followers, Total price, Price, Square, Living Room, Number of the Drawing room, Kitchen and Bathroom, Building Type, Construction time. Renovation condition, Building structure, Ladder ratio (which is the proportion between the number of residents on the same floor and the number of elevator of ladder. It describes how many ladders a resident have on average), Elevator (the number of elevators), Property rights for five years (It's related to China restricted the purchase of houses policy), Subway, District, and Community average price.

3 Objective

In this study, I will try to check if decisive features are making prominent contributions to the total second-hand housing price in Beijing. If there are any of these features, how much can they overwhelm other features when deciding the total price is another valuable question arises. When we decide to purchase property, normally, we would leverage between the total price and the preferable characteristics of the product due to the limited budget: a house located in a better place (e.g., nearby subway station) is likely to be more expensive than the one with the same condition but located far away from the subway station. With the limited budget to purchase a property, if our priority is to commute to work by subway each day without other requirements, then we would like to know how much the total price could be decreased if we were willing to give up other housing features such as large square number or more rooms in exchange for

commuting more conveniently. Meanwhile, to evaluate the complexity of predicting second-hand housing prices in Beijing, I will try several regression models to see if compatibility is acceptable. Regression models are relatively general and much easier to interpret than other elaborated models such as various neural networks. Regression would be more applicable, clearly showing the potential relationship amongst feature variables and telling us what predictors in a model are statistically significant and which are not. [5] Suppose the linear models have a sound performance. In that case, we could likely apply them to the post-pandemic second-hand housing price prediction in Beijing and other cities as well, as the reason mentioned in the introduction part. While neural networks could generate a more accurate prediction, another advantage of a regression model over a neural network is that the hardware requirement for training neural networks is higher than for training a regression model. Considering limited resources and time, if we want to conduct an efficient analysis on a huge dataset, then we are better off applying regression analysis over deep learning.

4 Methodology

Data pre-processing matters. In order to have a valid dataset to come out with a precise and accurate prediction for the future trading and prices of houses, we should first conduct pre-processing. Firstly, the URL and id columns were dropped. By definition, each second-hand house would be assigned a unique URL under the second-hand property selling platform path

(lianjia.com). Each of these unique URLs is linked to the information page where the related second-hand house feature values were posted.

Similarly, every second-hand house ready to be sold would be assigned a unique id to distinguish it from other second-hand houses on the platform. Those two variables have no meaning more than serving as house identifications, so I deleted them to diminish potential distraction for later prediction. Another two columns I chose to delete are the Lat column and the Lng column. Those two are the latitudes and longitude of houses in Beijing, providing the exact geographical location, which is highly correlated with the district number column (different districts of Beijing would be assigned with a unique integer). Since the number of unique values in these columns is much larger than the number in most of the other variable columns, I decided to delete those two and keep the district column as a replacement (Besides, I generated a correlation map between each of these feature column and the total price column Fig. 1, which also shows that the correlation of Lat-totalPrice and Lng-totalPrice are trivial).

Then, considering that some unknown characters are mixed with the floor number in the floor column, I extracted the actual floor of each observation. tradeTime column contains the date at which the house was sold successfully. Meanwhile, the range of this column begins from "2009-01-01" to "2018-12-31", which, similarly to the Lat column and Lng column, has too many unique values compared to most other columns. For the model to detect the trending efficiently, I extracted the year, month, and drop date to reform the tradeTime column. For those missing values (nan) and trivial values (some cells would consist of unknown characters), they were deleted considering that it would be extremely time-consuming during model training (if the model trained with the abridged dataset can generate a good prediction, then it suggests that the model can fit the data well. When it comes to the real production or business market, we

should apply the whole available dataset to achieve more accurate results.). More importantly, training models on the largest possible number of observations with completed feature columns would help to make the model more comprehensive and, thus, get better predictions afterward. Although the number of observations with abnormal feature values stated above is large and represents even more than half of the number of the whole observations (166585 out of 318,851) Fig. 2, the rest of the dataset is still huge enough for us to test whether or not our regression models could fit the second-hand housing price decently (with testing that conducted later, it did show that even with a much smaller training dataset could already generate a regression model fitting the price trending well.).

Lastly, I downsampled the dataset based on the number of houses traded during each unique trade time period I reformed. The property market would fluctuate yearly due to various factors such as deflation or periodical economic crisis. Thus, in which year a house is posted to the market can generate different price estimations. One of the goals of this study is to detect the second-hand house price trend. If the dataset is imbalanced on the tradeTime column, then the model trained after would concentrate on a specific period and spend much less effort to find the trending during other trade periods, which results in a model that is not representative enough to conduct house prices prediction on a long time scale. In Fig. 3, we can see that the dataset contains extremely imbalanced tradeTime values: Most of the houses were traded between the year 2016 and the year 2017. In order to have a balanced training dataset concerning trade time and further decrease the training dataset for more efficient checking towards the model's trending fitting performance as well, I randomly selected 20 observations out of each group consisting of all observations with the same tradeTime if the number of observations in that group is more than 100; For those groups that contain less than 100 observations, I included them all. We can

see that in Fig. 4, after downsampling, the dataset was much more balanced respective to tradeTime.

After data pre-processing, it is time to conduct data analysis with models. I tracked the trending of this dataset with three models: PLS, MLR, and RF. Before that, I loaded the abridged dataset into ProMV to generate Fig. 5 and Fig. 6, which are Square Prediction Error and Hotelling's T Square (denoted as SPE and HTS, respectively for short) of the processed dataset respectively. As we can see in those two figure, there are 7 and 17 observations out of total 1808 observations ranged outside the 99% confidence interval of HTS and SPE, respectively (Both less than 1% percent out of total number). Unlike analyzing industry data, where process monitoring is widely applied, and glitches such as false alarming (type I error) and false negative (type II) are likely to occur due to the aging of machines and negligence of operators, this dataset should have much fewer concerns over those outlier issues. The information of each house on the lianjia.com platform was provided and uploaded by each related individual property owner. They have extremely high confidence levels regarding their house information. When they post their property to be sold on the online platform, they must make sure that their house descriptions are consistent with reality. Thus, the number of potential outliers is highly under control (I regard the observation with SPE or HTS located outside the 99% confidence interval as potential outliers to be checked) and is limited within a reasonable range. Considering the situation stated above, those potential outliers are very likely to be the normal observation with correct feature values, except that they do not have common housing layout. In the end, besides the reasonable number of potential outliers, to keep predicting model inclusive for the largest possible types of second-hand houses, those observations were kept (In fact, after testing, the prediction performance

generated by model with removing potential outliers is almost the same as the one derived by keeping those potential outliers.).

After setting the total price column as the label and the rest of the variables as feature columns, I first trained PLS and MLR with the processed dataset. I set the validation number to 10, the random state as 1, and each validation would repeat three times for both models; for PLS, I also iterated the training and validation result generated by setting the number of components (loading vectors) from 1 to 17. After that, I re-fitted those two models with 70% of the processed dataset and tested with another 30% to derive R2, MSE, and RMSE. Then I trained RF with fine-tuning on 70% of the processed dataset and derived the best parameter with related metrics. Equipped with the best parameter, I re-fitted the model with 70% of the processed dataset and derived metrics with the rest of the dataset. The results are decent considering the abridged dataset's range, and I will explain more later.

5 Results and Discussion

As Fig. 1 shows, the top five (with correlation score with the total price larger than 0.4) most predominant features in deciding whether the total price is high or low are price (price per square), square, communityAverage, and the number of bathrooms (it is obvious that price is the just the total price divided by the square, which is highly correlated to the total price. If the price per sure is high, then there is no doubt that the total price will be high considering that most of houses to be sold have a certain range of square). This processed abridged dataset can represent the correlation map for the dataset with the same feature columns but before abridging well. This map suggests that those features with high correlation score are very likely to move together with

the total price in the same direction. With this trend, we could say that besides the price per square, the second-hand house price in Beijing heavily depends on the square, community average price, and the number of bathroom.

With the correlation map, we can deduct that if the feature we require has large correlation score, then with limited budget, we have to either give up a feature with larger correlation score or give up several features with smaller correlation scores to achieve our priority. We could apply this trend to conduct a quick analysis of second-hand house prices in Beijing when we are asked to provide consultancy.

After analyzing, models PLS, MLR, and SVR were trained with the dataset, and the metrics generated were very decent, and it showed that the second-hand house price could be interpreted really well by those regression models. Fig. 7 shows the R2, MSE, and RMSE validation curves with a different number of components after training the PLS model with 10-fold cross-validation (each validation repeated three times) and setting the random state to 1. As we can see, MSE and RMSE decrease as the number of components increases, and the trending is converse for R2 (In the following explanation, we apply MSE to represent the error decreasing.). The prediction error decreased prominently when I added 17 components one by one. By adding one more component, the MSE could be decreased by around half. But after adding the 8th component, the downward trend is trivial in Fig. 7 (MSE=0.16347 for 1 component, 0.08640 for two components, 0.04942 for three components, 0.02980 for four components, 0.01865 for five components, 0.01232 for six components,0.00829 for seven components, 0.00396 for eight components, 0.00134 for nine components, 0.00045 for ten components, 0.00014 for 11

components, 2.3518e-05 for 12 components, 1.1403e-06 for 13 components, 3.0758e-07 for 14 components, 7.9892e-08 for 15 components, 5.3822e-09 for 16 components, and 5.1298e-10 for 17 components). Similarly, for R2, after the 8-the component was added, the upward trend is trivial, shown in Fig. 7 (R2=0.83427, 0.91215, 0.95002, 0.96938, 0.98098, 0.98752, 0.99152, 0.99588, 0.99871, 0.99964, 0.99989, 0.99999, ~1 (0.9999994086586893), ~1 (0.9999999607761247), ~1 (0.9999999910030158), ~1 (0.9999999983798851), ~1 (0.9999999998138591) for adding the first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, eleventh, twelfth, thirteenth, fourteenth, fifteenth, sixteenth, and seventeenth component respectively.). After training with validation, I re-fitted the PLS model on 70% of the processed dataset with ten components. The derived metrics of testing on the rest of those 30% datasets are pretty positive: 0.99955 for R2 and 0.00033 for MSE.

When testing with MLR, I found that the metrics were even better than those produced by PLS. Training on the processed dataset with 10-fold cross-validation (each repeat three times) and setting the Radom state to 1, I derived that the validation MSE was around 0 (2.466210541318819e-30) and R2 was 1.0. The metrics after re-fitting MLR with 70% of the processed dataset and testing on the rest of the 30% dataset were just as perfect as validation metrics: around 0 for MSE (6.625689056916765e-31) and 1.0 for R2, which suggested that the trained MLR model explained 100% variance of the testing dataset, and it predicted the second-hand house price with filtered feature columns extremely well.

Lastly, for testing of RF mode, I first conducted fine-tuning on 70% of the processed dataset with 3-fold cross-validation. The parameter matrix was set to [15, 20, 25, 30, 35] for max_depth,

[120, 130, 140, 150] for n_estimators, and [6, 8, 10, 12] for max_features. After fine-tuning, I derived that the best parameters to choose within the certain range were {'max_depth': 25, 'max_features': 12, 'n_estimators': 140} for a validation score of 0.8806725233876023. After re-fitting the RF model on the same 70% of the processed dataset with the best parameters, I tested with the RF model on the rest of the 30% processed dataset. The metrics were still decent: 0.009807, 0.099028, and 0.987948 for MSE, RMSE, and R2, respectively. As we can compare, when we conduct prediction, the RF model would generate metrics similar to the PLS model we trained with six components would likely to generate (0.0086588523470536 and 0.9873350925056408 for MSE and R2, respectively.).

All three models I tested could perfectly fit the second-hand house dataset and provided a really accurate prediction on the total price based on given feature columns. Based on the metrics derived, MLR had the best performance over the other two. Surprisingly, the RF model, as one of the most popular machine learning models, was surpassed by PLS and MLR, which are two PCA methods.

Besides all, considering the interpretability, those three regression models above explain relationships in data, which other sophisticated methods are hard to achieve. So-called BlackBox models, such as neural networks, give us little information regarding their decision-making processes; the algebraic complexity of the functions they learn tends to lose any meaning with respect to the original set of feature variables. On the other hand, models that lend themselves to interpretability, such as linear regression (linear regression is good at interpreting: if x increases by 1, y would increase by a certain number, and we can all go home) and decision trees (what

decision tree does is splitting feature space into rectangles, ending up with a pretty diagram describing the logic behind the model) tend to fall short in the accuracy department, as they often fail to capture any nuanced or complicated relationships within a dataset. [6] Thus, to defy this interpretability-accuracy tradeoff, we can introduce MLR and RF, which are built based on linear regression and decision trees. Similarly, the procedure of PLS is trying to figure out the relationship amongst data features and, meanwhile, the relationship between the feature matrix and the label matrix/vector. Hence, we could try to apply those three regression models to conduct an analysis/prediction of Beijing's second-hand house price during pre- and, most likely, post-pandemic time, as we suggested already.

6 Conclusion and Future Work

This study was conducted to find out the potential trend behind Beijing's second-house market. This study successfully fitted three regression models with the potential trend, achieving really good outcome metrics. Beside, the correlation between each house feature and the total price was unveiled, helping to make future analysis and prediction for housing prices.

In the future, more work could be done to make this study more meaningful and comprehensive.

For methodology side, we could try other methods to deal with data imbalance such as upsampling. For downsampling applied in this study, it was conducted based on the tradeTime feature column. We could try to downsample based on each of those features to see which feature column would be the best to choose for the largest possible number of balanced feature columns

afterward, while maintaining the balance of tradeTime column. Besides, we could separate different housing types to train our models to deal with related prediction task so that those models will be more targeted and accurate.

References

[1] Michael, B. (2021). Bubble Economics: How Big a Shock to China's Real Estate Sector Will Throw the Country into Recession, and Why Does It Matter? International Journal of Housing Markets and Analysis., 14(5), 1111–1128.

[2] Zeng, S., & Yi, C. (2022). Impact of the COVID-19 pandemic on the housing market at the epicenter of the outbreak in China. *SN Business & Economics*, *2*(6), 53. https://doi-org.libaccess.lib.mcmaster.ca/10.1007/s43546-022-00225-2

[3] Li, S., & Lin, S. (2023). Housing property tax, economic growth, and intergenerational welfare: The case of China. *International Review of Economics & Finance.*, *83*, 233–251. https://doi.org/10.1016/j.iref.2022.07.010

[4] Ruiqurm. *Housing price in Beijing, Housing price of Beijing from 2011 to 2017, fetching from Lianjia.com. 2018*. CC BY-NC-SA 4.0. lianjia.com. Web.

[5] Terence Shin. (2021, Aug 10). *3 Reasons Why You Should Use Linear Regression Models Instead of Neural Networks.* https://towardsdatascience.com/3-reasons-why-you-should-use-linear-regression-models-instead-of-neural-networks-16820319d644

[6] Tom Grigg. (2019, Apr 8). *Interpretability and Random Forests.* https://towardsdatascience.com/interpretability-and-random-forests-4fe13a79ae34