Comparison of prices depending on factors in the secondary housing market

Vladislav Koktashev[1], *Vladimir* Makeev[1], *Pavel* Peresunko[1], *Anton* Mikhalev[1,*] and *Vadim* Tynchenko[1,2]

[1]Siberian Federal University, 79, Svobodny Av., Krasnoyarsk, 660041, Russian Federation
[2]Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., 660037 Krasnoyarsk, Russian Federation

Abstract. The article considers the issue of pricing for the secondary real estate market regarding local causes (physical properties of housing). The aim of the study was to verify the following hypotheses: the influence of the pricing factors of residential real estate on its value is determined by its price segment and the influence of infrastructure on the value of apartments in different cities is the same. In the hypothesis test, data were used on the secondary housing market of the cities of Novosibirsk and Krasnoyarsk, taken from the site of «CIAN» apartment sale announcements and from various open data sources. During the study, non-parametric methods of machine learning, model-agnostic methods for the interpretation of predictive models, hierarchical clustering are involved. As a result of the work, the first hypothesis was confirmed and the second hypothesis was refuted, the high accuracy of forecasting the cost of an apartment was achieved, and the peculiarities of price formation for secondary housing objects were revealed and described.

1 Introduction

The real estate market is one of the most important indicators of the development of the city. In this regard, the question arises of the pricing of real estate. The reasons affecting price formation can be divided into two groups: local and global [1-3]. Local causes are those reasons that create the entire palette of real estate prices at a given time and depend on the physical properties of the housing. Global factors are understood as macroeconomic parameters: the level of development of the city, the level of income and life of the population. In this work, there is a study of the pricing of real estate with respect to local causes.

When choosing an apartment in the high price segment of real estate, higher priority will be given to factors that differ from the factors that guide the choice of comfort class real estate [2]. In this regard, there is a suspicion that the price segment determines the influence of the pricing factors of residential real estate on its value.

The global reasons for the pricing of apartments make it possible to compare the general price level in different cities and state that the price of a similar apartment in different cities will be proportional to the general price level of the city [1]. Accordingly, the question arises whether local causes will affect the cost of apartments in different cities equally?

Work [3] presents a study of the impact of the infrastructure area in which the apartment is located when predicting its value. Based on the obtained models, the authors came to the conclusion that the factors of the location of the apartment have a significant effect on its value. Also, due to the large errors of models based on training data, they put forward the assumption that there are special pricing laws for this type of apartment.

However, a massive assessment of real estate through the direct use of functional data approximation can lead to difficulties in interpreting the found dependencies, inconsistency in estimates of their coefficients [4-6]. The author notes the appropriateness of applying the approaches considered by him only in combination with other methods of processing pricing relationships. For example, one of such methods may be factor analysis and clustering of real estate, followed by setting the response function for each cluster [7].

Particularly noteworthy are the methods of forecasting the cost of an apartment, taking into account spatial heterogeneity and the interdependence of real estate objects. Instead of describing the process of pricing the real estate market of the city using one model, these methods provide for the construction of linear models with variable coefficients, the values of which are determined by the site to which the property being evaluated belongs [8, 9].

2 Materials and Methods

The object of this study is the secondary real estate market, the subject of the study is the influence of pricing factors on the value of real estate indicated in the announcement of its sale.

* Corresponding author: spaming1@yandex.ru

The aim of the study was to verify the following hypotheses:

- The influence of the pricing factors of residential real estate on its value is determined by its price segment.
- The impact of infrastructure on the cost of apartments in different cities is the same [10].

The data set used is structurally completely identical to the data set collected for the city of Krasnoyarsk in the work [3]. The main sources of information: a register of open data from the Fund for Assistance to the Reform of Housing and Communal Services, a database of announcements for renting and selling real estate of CIAN, as well as data from the Yandex.Directory service. In order to improve the quality of the data collected, the ads of real estate agencies and realtors were removed from the sample. For this, a restriction has been introduced on the number of repetitions of the telephone number in the received sample - no more than 3 announcements per one number. The final sample consists of the following parameters: number of rooms, total area, floor, maximum number of storeys of the house, type of walls and ceilings, repairs, number of balconies and loggias, type of bathroom, number of freight and passenger elevators, year of construction of the house, type of house heating, house breakdown as well as geolocation features. The number of organizations of various types located within a radius of 1000 meters from the house was used as geolocation features. The types of considered organizations are: bank, bar, kindergarten, gas station, cafe, cinema, metro station, clothing store, public transport stop, park and public garden, clinic, post office, dentistry, grocery store, moll, university, school. The total sample size was 2765 records [10].

During the study, nonparametric machine learning methods are involved:

- An insulating forest for detecting abnormal observations, which is the most suitable way to search for emissions compared to the probabilistic approach and metric methods [10].
- Random forest and gradient boosting of decision trees to predict the value of the property.

Quality metric of regression models - mean absolute percentage error (mean absolute percentage error - MAPE): for random forest models, its value was determined by the predictions on unselected samples, for gradient boosting models - according to the forecasts obtained in the process of moving control with 10 blocks and 5 repetitions.

The analysis and interpretation of regression models was carried out by model-agnostic methods Permutation Feature Importance, Accumulated Local Effects Plot, set out in [11].

Assessment of the degree of heterogeneity of the data was carried out according to Hopkins statistics: calculated statistics values exceeding 0.75 correspond to the rejection of the null hypothesis of a uniform distribution of data in favor of the alternative with a significance level of 10% [12]. The search for groups of nearby objects was carried out by the method of agglomerative hierarchical clustering for the matrix of Euclidean distances between observations, the estimation of inter-cluster distances was carried out according to the Ward's method [3].

3 Results and Discussion

Anomalies were detected in the source data using the isolation forest method [10]. For each observation, an anomaly estimate was obtained, after which 48 elements were excluded from the sample for which the value of this criterion exceeds 0.54.

A model of a random forest is constructed to predict the cost of apartments in Novosibirsk. The value of the MAPE model on the selected samples – out-of-bag error, estimating the magnitude of the error in the total population of objects - amounted to 11% [13]. Out-of-bag error of a similar model obtained on the initial sample without geospatial features was equal to 15%.

Figure 1 shows a diagram of the regression residuals, which clearly shows an increase in the spread of errors with an increase in the true value of the apartment.

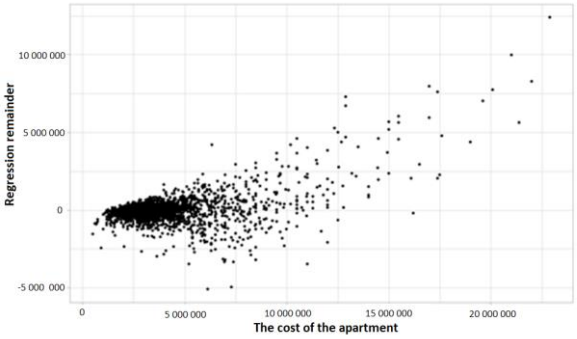

Fig. 1. Scatter chart of differences in forecasted and observed values.

In order to identify homogeneous groups of sample objects, the most important features in the model are identified and selected using Permutation Feature Importance algorithm (Figure 2, 10 most significant signs) [11]. A subsample has been generated for clustering of features that have a significance measure value greater than 1.5. The value of Hopkins statistics for this sample is 0.89, which confirms the heterogeneous nature of the data noted on the graph of the remnants of the regression model, and the presence of a tendency to group them.

Grouping of subsampling elements using the hierarchical clustering method was performed. As a result, two groups of residential real estate objects were obtained: the first includes 2085 objects, the second - 632. The resulting groups are shown in Figure 3. The average values of the most different from each other signs, rounded to hundredths, are presented in table 1 [9].

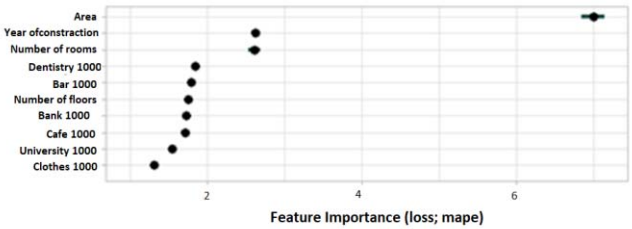

Fig. 2. The significance of the random forest model factors.

Table 1. The average values of the characteristics of the groups of real estate.

Characteristics	The average value of the characteristics	
	First group	Second group
Metrostation 1000	0.11	1.52
Cafe 1000	1.52	20.39
Bar 1000	1.94	15.24
University 1000	2.74	18.82
Cinema 1000	0.35	2.16
Clothes 1000	5.08	29.04
Bank 1000	5.31	30.17
Dentistry 1000	6.51	26.65
Park 1000	0.89	3.09
Apartmentboiler	0.01	0.03
The cost of the apartment	3563199.95	5368649.54

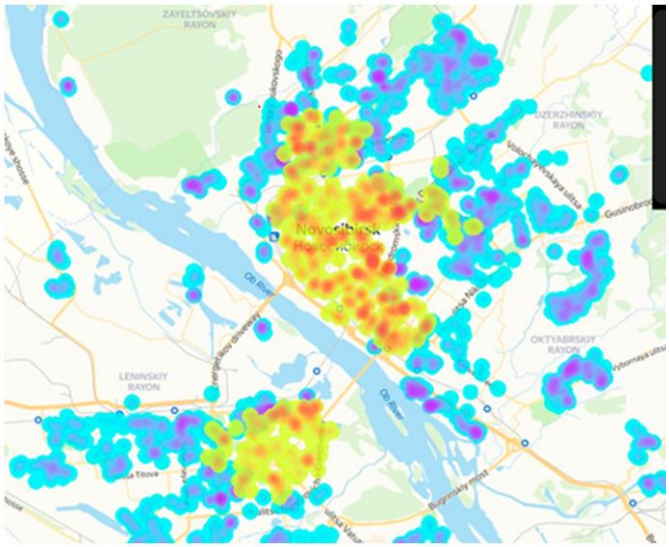

Fig. 3. Heat map of the cost per square meter of residential real estate.

For each of the selected groups of objects, a gradient boosting model is constructed that predicts the value of residential real estate (MAPE in both cases was 11%). The effect of the physical characteristics of the house and apartment, which is a subset of the most important distinguished features (see Figure 3), on the average output of the models is investigated. The effects were calculated using the Accumulated Local Effects Plot method; the corresponding graphs are presented in Figure 4 [4].

For the initial sample data on residential real estate in Novosibirsk and a similar sample of the city of Krasnoyarsk described in [3], common features are identified, random forest models are constructed to predict the cost of an apartment with the same hyperparameters. The accuracy of the models did not change: the MAPE values remained at the same level as was observed for models trained on data with a full set of features. The importance of each factor of both models is determined using the method previously used to select traits for clustering data. The obtained values of measures of the importance of features characterizing the infrastructure of the area around the apartment are displayed on the graph shown in Figure 5 [17].

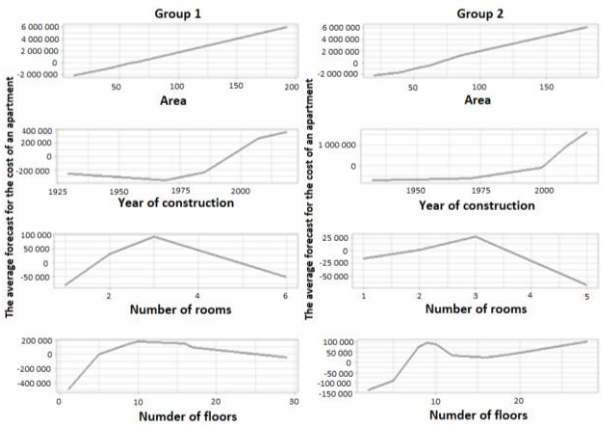

Fig. 4. The influence of signs on the average output of models.

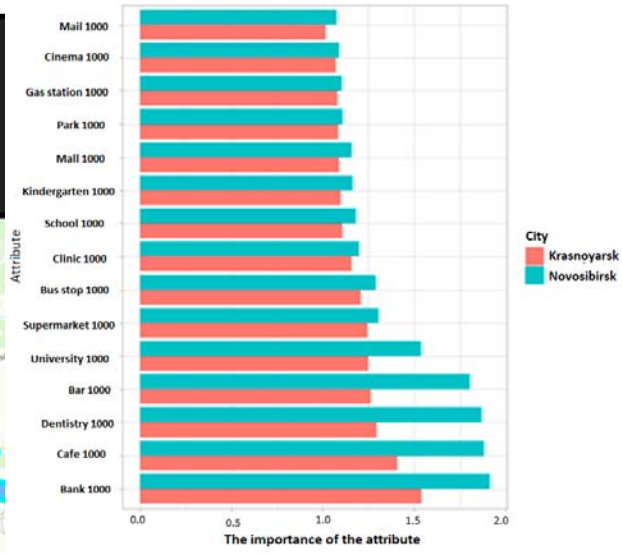

Fig. 5. The importance of geospatial features.

The models of a random forest predicting the cost of apartments in Novosibirsk have the following features:

- As in the case of Krasnoyarsk [3], for Novosibirsk there is also the problem of assessing the value of apartments of a high price category.
- The value of the MAPE model was 11%, which is less than the error characteristic of the models for mass valuation of real estate [14].
- The effect of improving the quality of the model due to the inclusion of infrastructure features observed in [3] for Krasnoyarsk is also present in Novosibirsk.

The growth of deviations between the predicted and the true cost of apartments in Novosibirsk is more directed towards underestimating the present values of the predicted model feature.

- Only a small number of features have a significant effect on model output. Among them, there are quite obvious pricing factors (apartment area, year of construction of the house, number of rooms, number of storeys of the house), and geospatial signs, which is explained by the fact that the signs that have the most significant influence on the output of the model are described by organizations more represented in areas with well-developed infrastructure, such as the city center. Groups of signs of wall materials, house ceilings, and the number of bathrooms practically do not contribute to the output of the model: the secondary housing market is regulated by demand, and not by the costs of construction production [15].

The following features are characteristic of the groups obtained by clustering the data of the Novosibirsk real estate market and the corresponding gradient boosting models.

- An increase in the average values of the area and cost of an apartment by groups is also accompanied by an increase in the average cost per square meter of housing: for the first group, it is 65,582 rubles, for the second - 84,160 rubles.
- Between the area and the predicted value of the apartment for both groups, a linear type of connection is noted, for other signs it is more complex.
- The influence of the year of construction on the average output of the model in the second group is not constant: it grows positively with an increase in the values of this attribute, which cannot be said about a similar dependence for the first group.
- The influence of the number of storeys on the average model output for apartments located in houses with more than 15 floors is opposite in each group: for the first group, it consists in decreasing the predicted value with an increase in the number of storeys in the house, for the second - in increasing it.
- The location of residential real estate significantly affects its value. Apartments from the second group, having a higher cost per square meter, are located mainly in the city center and surrounding areas, while apartments from the first group are located mainly on the outskirts of the city.

In general, the influence of infrastructure on the cost of an apartment in Novosibirsk is slightly higher: the largest difference between the significance is typical of those pairs of geospatial features of random forest models that describe the number of banks, cafes, dentistry, bars and universities in a radius of 1000 meters from the house in which the apartment is located [16].

5 Conclusions

The results obtained give reasons to confirm the first and refute the second hypothesis, namely:

- The influence of pricing factors on the value of residential real estate in different price groups may have a heterogeneous quantitative nature and an uneven trend.
- The influence of infrastructure factors on the formation of housing prices for different cities is different: perhaps one of the reasons for this is the variety of conditions for the formation of the urban environment. However, the results show that in both cases, the signs of infrastructure positively affect the quality of the forecast model for the cost of apartments.

Due to the rather high accuracy achieved in forecasting the cost of apartments and the visual clarity of the graphs given when interpreting machine learning models, the approaches used in this study can be applied to constructing models for mass valuation of real estate and analyzing the real estate market from a global perspective [17].

References

1. V.V. Koktashev, V.V. Makeev, E.G. Shchepin, P.V. Peresunko, *Jour. of Phys.: Conf. Ser.* **1353**, 012139 (2019)
2. F.T. Liu, K.M. Ting, Z.H. Zhou, *ACM - TKDD.* **6**, 1 (2012)
3. A.A. Boyko, V.V. Kukartsev, V.S. Tynchenko, V.A. Kukartsev, E.A. Chzhan, A.S. Mikhalev, J*our. of Phys.: Conf. Ser.* **1333**, 032009 (2019)
4. N.V. Fedorova, V.V. Kukartsev, V.S. Tynchenko, Y.V. Danilchenko, S.N. Ezhemanskaya, N.V. Sokolovskiy, *IOP Conf. Ser.: Mat. Sci. and Eng.* **734**, 012084 (2020)
5. Z. Gongfu, *3rd Int. Conf. on Inf. Man., Inn. Man. and Ind. Eng*. **1**, 113 (2010)
6. N.V. Fedorova, V.V. Kukartsev, V.S. Tynchenko, Y.V. Danilchenko, S.N. Ezhemanskaya, N.V. Sokolovskiy, *IOP Conf. Ser.: Mat. Sci. and Eng.* **734**, 0120834 (2020)
7. S.Y. Kalashnikov, A.E. Godenko, Y.S. Kalashnikova, I.A. Tarasova, I*OP Conf. Ser.: Mat. Sci. and Eng.* **698**, 066003 (2019)
8. C. Zhou, *Jour. of Phys.: Conf. Ser.* **1802**, 032034 (2021)
9. M. Itoh, Y. Hagimori, K. Nonaka, K. Sekiguchi, *Jour. of Phys.: Conf. Ser.* **744**, 012222 (2016)
10. S.Y. Kalashnikov, A.E. Godenko, Y.S. Kalashnikova, I.A. Tarasova, *IOP Conf. Ser.: Mat. Sci. and Eng.* **698**, 066003 (2019)
11. L. Maksimenko, O. Korobova, O. Dudinova, X. Soskova, *IOP Conf. Ser.: Mat. Sci. and Eng.* **953**, 012043 (2020)
12. I.A. Panfilov, Y.A. Alekseeva, T.A. Panfilova, O.I. Karelin, D.V. Kustov, *Jour. of Phys.: Conf. Ser.* **1399**, 055092 (2019)

13. X. Liu, Q. Zhou, C.A. Wang, *Jour. of Phys.: Conf. Ser.* **1187**, 052104 (2019)
14. N. Lavrov, A. Druzhinin, N. Alekseeva, *IOP Conf. Ser.: Mat. Sci. and Eng*. **940 (1),** 012042, (2020)
15. O. Grineva, I. Shirochenskaya, *Jour. of Phys.: Conf. Ser.* **1425**, 012069 (2019)
16. F. Cui, *Jour. of Phys.: Conf. Ser.* **1629**, 012071 (2020)
17. Z. Tarmidi, N.H.A. Maimun, *IOP Conf. Ser.: Earth and Env. Sci.* **540**, 012047 (2020)