Pricing modeling in the housing market with the urban infrastructure effect

V.A. Koktashev[1], V.E. Makeev[1], E.S. Shchepin[1], P.V. Peresunko[1] and V V Tynchenko[1]

[1] Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russian Federation

E-mail: 051301@mail.ru, vkoktashev-ki16@stud.sfu-kras.ru, vmakeev-ki16@stud.sfu-kras.ru, Eshchepin-ki16@stud.sfu-kras.ru, pvperesunko@gmail.com

Abstract. At present, in many large industrial enterprises, one of the motivational incentives for hiring employees, as well as for encouraging them, is the provision of temporary or permanent use of residential real estate. This raises the question of choosing the best option of the property from a variety of proposals on the market. This article addresses the issue of estimating the value of residential property in the city of Krasnoyarsk. The descriptive signs of the apartment were not only its internal parameters, such as the area or number of rooms, but also the external characteristics that describe the environment of the apartment house. The data on the apartments were taken from the website of the apartment sales announcements and from various open data sources. The number of organizations of each type considered within a radius of 1000 m serves as a quantitative measure of the house environment. The model built using a random forest shows good results and solves the problem. The relative error of the forecast is 8%. In addition, it is shown the positive impact of the apartments external characteristics on the quality of the constructed model. As a result, an easily scalable model was built that can be applied to other cities.

1. Introduction

At present, in many large industrial enterprises, one of the motivational incentives for hiring employees, as well as for encouraging them, is the provision of temporary or permanent use of residential real estate. This raises the question of choosing the best option of the property from a variety of proposals on the market. Real estate has many different parameters that affect its value. The process of pricing itself is complicated, so it is not always possible to understand the optimal cost of an apartment. Knowledge of the apartment approximate cost will speed up the sale and reduce costs when buying residential real estate.

To date, proposed several ways of estimating the value of the apartment. For example, in the article [1], the authors set a goal to compare the accuracy of the hedonic model prediction using an artificial neural network. The authors randomly selected 200 homes in Christchurch, New Zealand. For building models, factors such as the size of the house, the age of the house, the type of house, the number of bedrooms, the number of bathrooms, the number of garages, amenities around the house and

geographical location were taken into account. The hedonic model includes a regression of observed house prices with those attributes of the house that are presumably decisive for the asking price. In this article, the authors used a semi-log model, since the price is a very sensitive and volatile component [2-4]. The model of the neural network is similar to the process used in the construction of the hedonic model [5-8]. To determine the optimal model of an artificial neural network, the trial and error method is used [9-11]. As a result, the neural network model was better than the hedonistic model. However, both models showed that variable locations play an important role in housing prices.

Unlike work [1], article [3] investigated the linear relationship between location and rental value of property. The authors conducted their research, examining the apartments of the city of Portland, Oregon. Before building the model, the authors conducted 670 observations of apartments. The results show that as the distance from the city center, within 10 miles, the rental cost falls. However, then the rental price for 7 miles increases. This is due to the relocation of residents in the suburbs, for the ring road around the city center. Then, as the apartments move away from the ring road and the city center, the rental value again decreases. According to the collected data on apartments, the researchers built a model:

$$R_i = \beta X_i + \gamma_1 DCC_i + \gamma_2 DCC_i^2 + \gamma_3 \mathrm{DCC}_i^3 + \gamma_4 \mathrm{DH}_i + \gamma_5 \mathrm{DH}_i^2 + \gamma_6 \mathrm{DI}_i + \gamma_7 \mathrm{DI}_i^2, \quad (1)$$

where $\mathrm{R_i}$ – monthly apartment rent, $\mathrm{X_i}$ – vector of apartment attributes, β – vector of implicit price limits for these apartment attributes, $\mathrm{DCC_i}$ – distance from the city center to the apartment, $\mathrm{DH_i}$ – distance from the nearest highway to the apartment, $\mathrm{DI_i}$ – distance from the nearest intersection of two highways to the apartment, e – stochastic mistake. The resulting model is well described the relationship between distance from the center and the rental value of the apartment in the city of Portland

It is also possible to predict the price of real estate value using various methods of forecasting. For example, in the articles [8–15], the authors use such algorithms as logistic regression, SVM, Lasso regression, decision tree regression, random forest regression, and neural networks. The selling price of real estate for these studies was determined by such factors as the location of the house, the material of the house, the area of the apartment, the age of the house, the number of rooms, etc.

Thus, the apartment cost is influenced not only by its internal parameters, such as the area or number of rooms, but also external parameters that describe the apartment location.

2. Materials and methods

2.1. Problem statement

It is necessary to build a forecast model for the estimated cost of the apartment in Krasnoyarsk city. The data for setting up the models are ads for the sale of an apartment on the CIAN website. Therefore, it is required to build a model to determine the estimated cost of an apartment that people put in the ad. Knowing this price, the company will be able to determine whether the apartment is not overvalued.

2.2. Data

For the study, data were taken from three sources: the database for renting and selling real estate CIAN (information about prices and characteristics of individual apartments), the register of open data of the Housing and Utilities Reform Facilitation Fund, and Yandex.Spravochnik website data. In order to increase the quality of the data obtained, it was decided not to consider ads from real estate agencies. For this purpose, a restriction on the number of repetitions of a phone number in the received sample was introduced - no more than 3 announcements per number. It was also introduced a limit on the cost of an apartment - no more than 4 million rubles. This restriction is justified by the assumption that the pricing of more expensive apartments are subject to slightly different rules and dependencies. Parameters considered: number of rooms, studio, total area, floor, number of floors in the house, parking, type of repair, number of balconies, number of balconies, type of bathroom, number of

elevators, garbage disposal, year of construction of the house, accident rate of the house, type of floors, wall type, the number of different types of organizations in radii (100, 200, 300, 500, 1000 meters). The total size of the resulting database is 1970 records.

As input geolocation signs of the model, signs were selected that describe the number of a certain type of organizations within a radius of 1000 meters. For example, the number of universities or bars in a radius of 1000 meters around the apartment. 15 types of such organizations were selected.

2.3. Methods

The feature selection was performed using the method of recursive elimination of features, where the solution of the regression problem was carried out using the random forest machine learning algorithm, and the ridge and linear regression methods. The criterion for choosing the best model was RMSE (root of the mean square error) at the cross-validation, determined by the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2}{n}}, \quad (2)$$

where y_{ij} - value of the response at the j-th point of the validation set of the i-th model; $\hat{y}_{ij}$ - output of the i-th model at the j-th point; k - number of cross-check blocks (10 blocks); m - number of elements of the validation sample; n - size of the original sample.

After selecting the optimal hyperparameters with a grid search, the MAE (mean absolute error) value of elementwise cross-check (LOOCV (Leave one out CV)) was calculated to determine the error in the same measurements (in rubles):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (3)$$

where y_i- response value of the control dataset for the i-th model; $\hat{y}_i$- output of the i-th model at the control sample point; n- size of the original sample.

3. Results and discussion

Errors of models with selected parameters are shown in Table 1.

Table 1. Models errors.

Model name	MAE
Random forest	209143
Ridge regression	259029
Linear regression	252884

Models created by means of the “random forest” algorithm have the lowest mean absolute error. Figure 1 shows the result of the model.

All apartments were ordered in ascending order of their value. Red dots were shown forecast model. The forecast was obtained for the elements of the original set that were not included by the bootstrap method in the training sample of one or another decision tree (out-of-bag sample).

An assessment of the importance of the factors of the best model. To assess the importance of using the approach based on accuracy (Accuracy-based importance). The importance of the most influential signs are contained in Table 2.

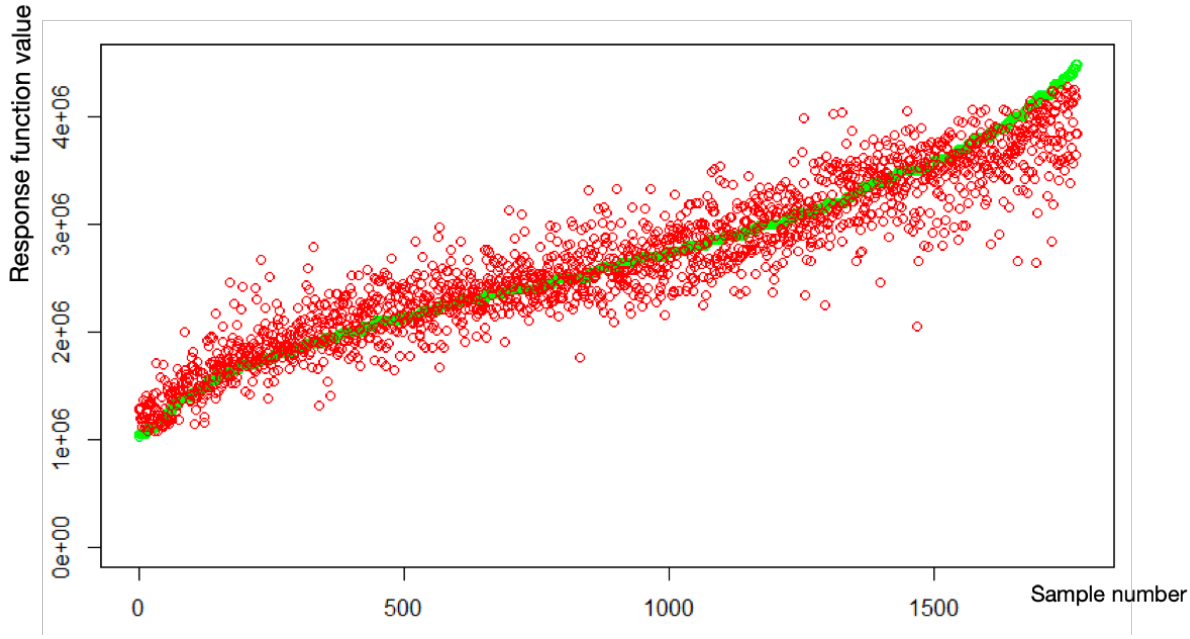

Figure 1. The result of the model, where: green points - true prices; red points - predicted prices.

Table 2. Error differences by feature.

Feature	Difference of RMS Errors
Number of rooms	111996315753
Total area	716580607833
Floors in the house	43803859690
Year of construction	54685779085
University 1000	17740672694
Dentistry 1000	28360210659
Clothes 1000	15668792050
Cafe 1000	25122370463
Kindergarten 1000	14601106645
Bar 1000	47876893832

It is noteworthy that six of the parameters contained in the table quantitatively describe the infrastructure of the area in which the apartment is located.

A comparison was made of the optimal of the above models and a model constructed in a similar way, the initial sample of which does not contain data on the location of the property. The MAE value for the latter was 264974 rubles. For the ratio of the models, the values of their mean absolute percentage error (MAPE) forecast for "out-of-bag" sampling were calculated using the formula:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}, \quad (4)$$

where y_i- predicted value in the i-th point of the sample; $\hat{y}_i$- model output to the i-th data set point; n - volume of the "out-of-bag" set.

The error of the model, including the location parameters, turned out to be 8%; the model without those - 10% (values are rounded to tenths). Thus, geolocation signs improve the quality of the model.

Of the three models, the random forest proved to be the best on the forecast. The forecast error of this model was 209,143 rubles. This is a relatively large amount, but it is important to understand that this is the cost that people put in the ad. The cost of the property, contained in the announcement of its sale, is not final. Due to the subjective nature of the value of the property declared by its owner, the cost of apartments with the same (close) characteristics may differ significantly. Therefore, the model itself is influenced by this subjectivity, therefore such an error.

In some cases, between the buyer and the seller may negotiate the nature of the transaction, including the cost of the apartment. Therefore, the predicted value of the response in the sampling sets used in the creation of statistical models may differ from the true one - what will (was) indicated in the purchase and sale agreement. This problem can be solved by using data on already completed purchase and sale of residential real estate. However, in this paper we do not pursue this goal, since it is important for us to predict the value that people put on ads.

It should be noted that the developed system is quite scalable and to include in it the support of a city, it is only necessary to collect the necessary data set of a certain format, which is not a difficult task and can be solved using the already mentioned data sources. The data on large values of errors correspond to the models for which the samples were used, where the cost of an apartment varies from 499,000 to 1,5500,000 rubles. When creating models, the error values of which are contained in this work, data were used in which the value of real estate is in the range from 10,000,000 to 4,490,000 rubles. Therefore, the obtained models have insufficient predictive power with respect to data on high-priced apartments. Perhaps their inaccuracy is associated with the fact that there are special pricing laws for this type of property.

4. Conclusion

In this study, the model was built to predict the apartment price. As a result of the study, it was concluded that the location of residential real estate significantly affects its value. The applied method of feature selection noted the predictors belonging to this class as one of the most significant in all used data sets and for all types of considered models. Thus, the consideration of geolocation signs is an important aspect in conducting further research in this area, including the construction of models with greater accuracy in predicting the value of residential real estate.

References

[1] Limsombunc V, Gan C and Lee M 2009 House price prediction: hedonic price model vs. artificial neural network *American journal of applied sciences* **1(3)** 193–201

[2] Shonkwiler J S and Reynolds J E 2006 A note on the use of hedonic price models in the analysis of land prices at the urban fringe *Land economics* **62(1)** 58

[3] Frew J and Wilson B 2002 Estimating the connection between location and property value *Journal of real estate practice and education* **5** 17–26

[4] Shinde N and Gawande K 2018 Valuation of house prices using predictive techniques *International journal of advances in electronics and computer science* **5** 2393–835

[5] Bukhtoyarov V V, Tynchenko V S, Petrovsky E A, Kukartsev V V and Kuklina A I 2018 Evolutionary method for automated design of models of vortex flowmeters transformation function *Journal of Physics: Conference Series* **1118(1)** 012041

[6] Yu H and Wu J 2016 Real estate price prediction with regression and classification *CS 229 Autumn 2016 Project Final Report* 1–5. Retrieved from: http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf

[7] Milov A V, Tynchenko V S, Kukartsev V V, Tynchenko V V and Bukhtoyarov V V 2018 Use of artificial neural networks to correct non-standard errors of measuring instruments when creating integral joints *Journal of Physics: Conference Series* **1118(1)** 012037

[8] Pow N, Janulewicz Y and Liu L 2014 Applied machine learning project 4 prediction of real estate property prices in Montreal. *Retrieved from:* http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf
[9] Park B and Bae J K 2015 Using machine learning algorithms for housing price prediction: the case of Fairfax county, Virginia housing data *Expert systems with applications* **42.6** 2928-34
[10] Tynchenko V S, Tynchenko V V, Bukhtoyarov V V and Agafonov E D 2018 *RPC 2018 - Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications* 8482126
[11] Ruben D J 2002 Data mining: an empirical application in real estate valuation *FLAIRS Conference* 314-7
[12] Lim W T, Wang L, Wang Y and Chang Q 2016 Housing price prediction using neural networks *12th International conference on natural computation, fuzzy systems and knowledge discovery, IEEE* 518-22
[13] You Q, Pang R, Cao L and Luo J 2017 Image based appraisal of real estate properties *IEEE transactions on multimedia* **19(12)** 2751-9
[14] Tynchenko V S, Tynchenko V V, Bukhtoyarov V V, Tynchenko S V and Petrovskyi E A 2016 The multi-objective optimization of complex objects neural network models *Indian Journal of Science and Technology* **9(29)** 99467
[15] Huang Y 2019 Predicting home value in California, United States via machine learning modeling *Statistics, optimization & information computing* **7(1)** 66-74