Advances in Intelligent Systems Research, volume 174

Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Gradient Boosting–Based Machine Learning Methods in Real Estate Market Forecasting

Nikita Fedorov*
Institute of Information Technologies
Chelyabinsk State University
Chelyabinsk, Russia
fedorovni1996@yandex.ru

Yulia Petrichenko
Institute of Information Technologies
Chelyabinsk State University
Chelyabinsk, Russia
yulia-c@yandex.ru

***Abstract*—Several approaches can be used to estimate the value of residential real estate. The sales comparison approach requires assessing a number of comparable residential properties and determining the degree of their compatibility. The cost approach requires using the information on all construction costs. The income approach requires a large amount of data on the market capacity, operation cost, expected operating expenses, and competitive opportunities. For the end customer or buyer and often for appraisers and realtors as well, these methods would involve processing a considerable amount of information. The sales comparison approach is used more frequently, since sufficient data for other approaches might not be publicly available. Nevertheless, all these methods are quite complex, and using them to estimate the value of a residential property can be time-consuming if performed without automated valuation models (AVMs). In the paper, the method of data collection is described, and the analysis based on these data is carried out. Moreover, the housing affordability index is determined (shows the number of years required to purchase a residential property). Finally, the most appropriate forecasting method with the least error is chosen, and the parameters of residential properties are determined and ranked according to the degree of their impact on the price.**

Keywords*—real estate market analysis, forecasting, housing market, housing, real estate, residential real estate market value

I. Introduction

The 2019 Russian housing market was marked by two important events which can affect the residential property value. As of July 1, 2019, shared-equity construction was replaced by a new financing scheme that uses escrow accounts (bank accounts where all funds contributed are kept blocked until certain conditions are met) [1]. The standard value-added tax rate increase from 18% to 20% was another significant event [2]. All of these factors are already leading to the rise in the final price of residential properties for citizens.

The housing (mortgage) loan market rate has now reached its peak (16.823 billion rubles) in the entire modern history of the Russian Federation, and the weighted average rate has reached a minimum of 9.56%. The previous peak of 10.267 billion rubles was in 2014, while the weighted average rate was 12.2%. In 2015, the rate declined to 7291 billion rubles, and in the following years it increased [3]. Thus, the citizens became more interested in buying residential property, the household debt level, however, rose greatly and reached 6.2% of gross domestic product [4].

According to the analytical report from construction company BEL Development, the key interest rate can be cut, ranging from 5.5 to 6.5% per annum [5]. The demand will most probably continue to grow in the following years due to the fact that mortgages will become more affordable. The increase in the final housing price will be moderated by lowering the mortgage rate to 6–6.5%. This rate can maintain the real estate market demand from individuals, but in this case Russian economy risks entering the phase of stagnation in the nearest possible time when even national projects will not be able to increase the economic growth. Moreover, if new sanctions on Russia are imposed (concerning the banking sector and Russian sovereign debt), this can reduce individuals' capacity to pay which will affect the demand.

Proper valuation of residential real estate is essential for both appraisers (their salary and reputation depend on the quality of appraisals) and buyers as it is important not to overpay. Nevertheless, in order to assess a residential property, it is necessary to take into account its various characteristics [6, 7]. This process can be quite time-consuming and complex if carried out without automated valuation systems [8]. Such systems are based on neural networks [9, 10], or machine learning methods can be used (RIPPER, AdaBoost, SVM, Ridge, Random Forest) [11-15], which take into account characteristics of the building, the level of infrastructure development and even external factors. Forecasting the value with the aforementioned models instead of using a sales comparison approach, which involves finding approximately 5–7 similar residential properties in the market and calculating the final price, will permit appraisers to save time [16, 17]. These models may be useful for buyers who are not able to use the cost or sales comparison approaches to valuation: with their help, undervalued or overpriced properties can be found and bargain purchases can be made.

II. Data Collection

To create a model which will be able to estimate the final price of a residential property, the following data on buildings are needed: commissioning year, number of floors, playground nearby, energy efficiency class, type of the building, load-bearing walls, etc. There are no open data bases with such information about residential buildings, thus, it was necessary to parse (carry out a syntactic analysis of) the Dom.MinGKH website where main characteristics of buildings are given [18].

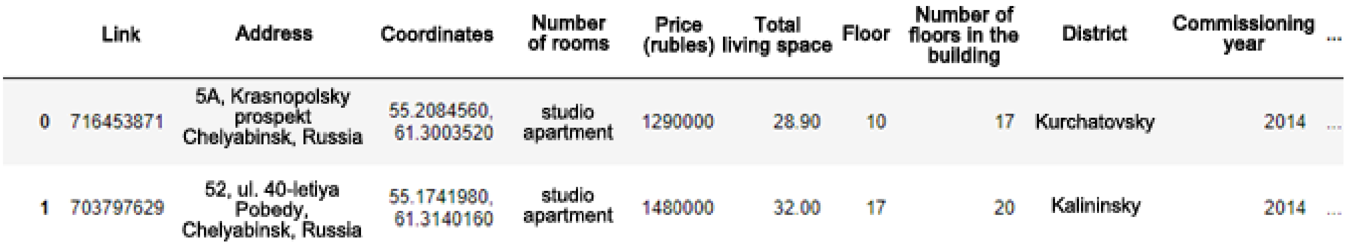
The data collection was carried out by using the following libraries: Beautiful Soup (HTML or XML files parsing), requests (sending requests to a server), csv (writing data in the right format with a convenient processing), and sleep (delay in sending requests) [19-22].

To forecast the price and identify important characteristics, it is necessary to obtain data on apartments for sale. In order to achieve this, it is required to parse a major real estate aggregator with a large number of listings and comprehensive information regarding apartment price, total living space, and floor number, however, such websites have a parsing protection which is difficult to bypass.

The DomClick website was chosen to collect data on residential properties [23]. There can be found various listings and data on all the necessary characteristics (price, total living space, commissioning year). The libraries mentioned above must be used, although, to bypass the protection, it is needed to not only make a delay for sending requests (the sleep library) but also change the User-Agent (a client application that uses a specific network protocol) because when sending requests via the library, the server may identify and reject the request or block an Internet Protocol address (IP address) if it sends too many requests. In order for requests to come from different IP addresses, it is important to change the API (Application Programming Interface) address and port number every time. The Free Proxy List website was used for this purpose [24].

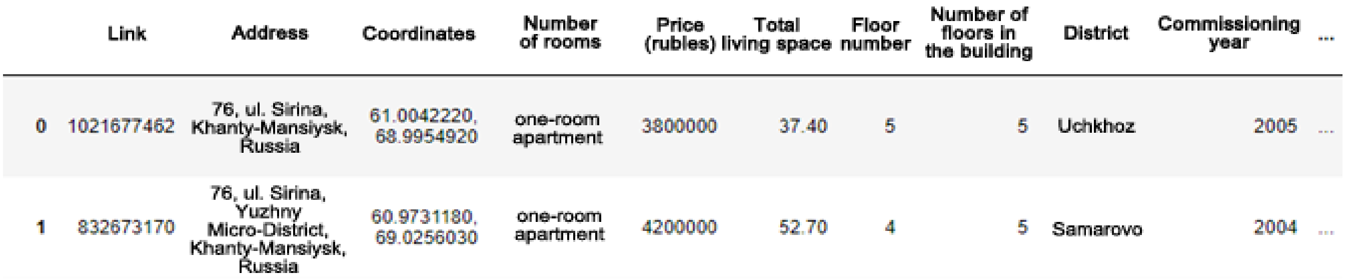
A single field which allows merging (address or coordinates) is needed to combine the received data. The Geocoder API by the technology company Yandex was used to obtain the address in a single format, however, this server does not provide data on the district which are needed for the further price forecasting. To determine the district in which the residential property is located, the OpenStreetMap service and the Nominatim library were used [25]. The project is non-profit and allows a limited number of requests. To remove the restriction, it is needed to change the User-Agent and IP address, as when accessing DomClick.

By combining the data using the VLOOKUPV function in Microsoft Excel, the data frame was obtained. It includes data on the residential property for sale and the characteristics of the building (shown in Fig. 1 and Fig. 2).



	Link	Address	Coordinates	Number of rooms	Price (rubles)	Total living space	Floor	Number of floors in the building	District	Commissioning year	...
0	716453871	5A, Krasnopolsky prospekt Chelyabinsk, Russia	55.2084560, 61.3003520	studio apartment	1290000	28.90	10	17	Kurchatovsky	2014	...
1	703797629	52, ul. 40-letiya Pobedy, Chelyabinsk, Russia	55.1741980, 61.3140160	studio apartment	1480000	32.00	17	20	Kalininsky	2014	...

Fig. 1. Example of the general population for Chelyabinsk.



	Link	Address	Coordinates	Number of rooms	Price (rubles)	Total living space	Floor number	Number of floors in the building	District	Commissioning year	...
0	1021677462	76, ul. Sirina, Khanty-Mansiysk, Russia	61.0042220, 68.9954920	one-room apartment	3800000	37.40	5	5	Uchkhoz	2005	...
1	832673170	76, ul. Sirina, Yuzhny Micro-District, Khanty-Mansiysk, Russia	60.9731180, 69.0256030	one-room apartment	4200000	52.70	4	5	Samarovo	2004	...

Fig. 2. Example of the general population for Khanty-Mansiysk.

III. Data Analysis

According to a survey conducted by the research company NAFI (The National Agency for Financial Studies), the main criterion citizens focus on when buying residential property is price [26, 27]. For the third consecutive year, it becomes the key criterion when one makes a purchase, followed by the location of the house and its planning. Thus, it is necessary to consider economic performance of the regions and determine the housing affordability index (HAI) which shows the ratio of the family's actual income to the ideal income needed to purchase a typical apartment [28]. In order to achieve this, the HAI formula should be used (1):

$$\text{HAI} = \frac{P \cdot S}{I \cdot M}, \tag{1}$$

where P is the average market price per square meter of living space, S is the total living space (a standard apartment with a total living space of 54 square meters), I is the per capita income, and M is the number of family members.

According to official statistics, the average gross monthly nominal wage in Khanty-Mansiysk is 73,780 rubles and the average price of 1 square meter of the total living space of apartments in the market is 52,276 rubles. As for Chelyabinsk, the corresponding figures are 36,487 rubles and 35,998 rubles [29, 30]. Depending on the level of the HAI, markets are classified according to the degree of housing affordability. International experience shows that the optimal index level should not exceed 3 years. The housing affordability index in Khanty-Mansiysk is at the level of 2.11 years, while in Chelyabinsk it is at 2.36 years. This means that if a family spends all their income to pay the mortgage, they will be able to repay it in 2 years (interest not included). On average, the index equals to 3.9 years in Russia, and Sevastopol has the highest rate – 11 years.

Furthermore, it is important to calculate how much time it can take for a family to repay the mortgage if they do not pay all the income but use some of the money for their needs. When granting a mortgage loan, it is considered that a household is able to make loan payments without loss if the payments do not exceed 30% of the income (2):

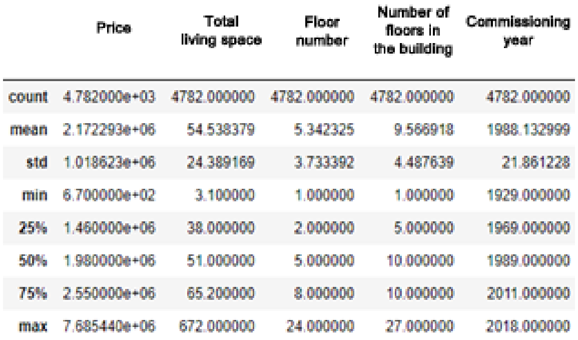
$$\text{HAI} = \frac{P \cdot S}{(I - 70\%) \cdot M}, \qquad (2)$$

where *P* is the average market price per square meter of living space, *S* is the total living space (a standard apartment with a total living space of 54 square meters), *I* is the per capita income, and *M* is the number of family members.

In Khanty-Mansiysk, a family will buy an apartment in 7 years if they pay 30% of the income, and in Chelyabinsk – in 7.9 years under the same conditions. These rates correlate directly to the demand in the real estate market, since the higher the housing affordability index, the lower the level of demand.

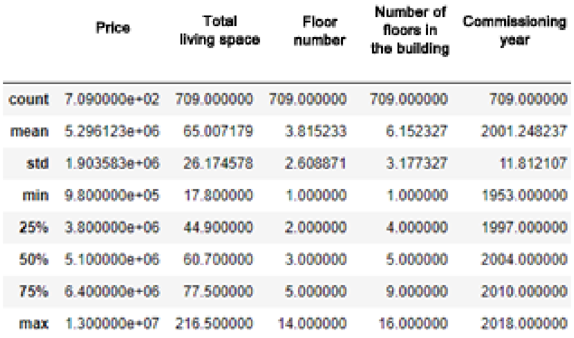
It is necessary to study main statistical characteristics for each attribute (the number of non-missing values, root-mean-square deviation, standard deviation, median, the first and the third quartiles) in order to determine the distribution of data. To perform this task, the describe function in Python was used.

As for Chelyabinsk, the average commissioning year of the buildings is 1988, the average number of floors is 5, and the average price of an apartment is 2.1 million rubles (the data are shown in Fig. 3). Moreover, a third of all apartments cost less than 2.5 million rubles.



	Price	Total living space	Floor number	Number of floors in the building	Commissioning year
count	4.782000e+03	4782.000000	4782.000000	4782.000000	4782.000000
mean	2.172293e+06	54.538379	5.342325	9.566918	1988.132999
std	1.018623e+06	24.389169	3.733392	4.487639	21.861228
min	6.700000e+02	3.100000	1.000000	1.000000	1929.000000
25%	1.460000e+06	38.000000	2.000000	5.000000	1969.000000
50%	1.980000e+06	51.000000	5.000000	10.000000	1989.000000
75%	2.550000e+06	65.200000	8.000000	10.000000	2011.000000
max	7.685440e+06	672.000000	24.000000	27.000000	2018.000000

Fig. 3. Main statistical characteristics for Chelyabinsk.



	Price	Total living space	Floor number	Number of floors in the building	Commissioning year
count	7.090000e+02	709.000000	709.000000	709.000000	709.000000
mean	5.296123e+06	65.007179	3.815233	6.152327	2001.248237
std	1.903583e+06	26.174578	2.608871	3.177327	11.812107
min	9.800000e+05	17.800000	1.000000	1.000000	1953.000000
25%	3.800000e+06	44.900000	2.000000	4.000000	1997.000000
50%	5.100000e+06	60.700000	3.000000	5.000000	2004.000000
75%	6.400000e+06	77.500000	5.000000	9.000000	2010.000000
max	1.300000e+07	216.500000	14.000000	16.000000	2018.000000

Fig. 4. Main statistical characteristics for Khanty-Mansiysk.

In Khanty-Mansiysk, the average commissioning year of the buildings is 2001 (as can been seen in Fig. 4). Therefore, the residential stock for sale in Khanty-Mansiysk is newer than the one in Chelyabinsk (1988).

The average price of an apartment in Khanty-Mansiysk is 5.3 million rubles and the average number of floors is 5.

IV. Price Forecast

To forecast the price, the four following models in Python were chosen:

- LinearRegression;
- CatBoostRegressor;
- CatBoostRegressor;
- AdaBoostRegressor.

LinearRegression (linear regression) is the simplest and most traditional regression model (3):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,, \qquad (3)$$

where β_0 is the *Y* intercept, β_1 is the slope of *Y*, and ε_i is the random error in *Y* for observation *i*.

Linear regression finds the characteristics which minimize the mean squared error between the predicted and the actual values. Mean squared error (MSE) is equal to the sum of the squared difference between the predicted and the actual values [31].

CatBoostRegressor is the first machine learning algorithm developed by Yandex and released as an open-source under the free Apache License 2.0. The algorithm is based on MatrixNet, a proprietary machine learning algorithm created by Yandex for using gradient boosting in the company's projects [32, 33].

XGBRegressor is a model based on the gradient boosting decision tree algorithm. In each iteration, prediction deviations of the previously trained ensemble in the training set are calculated. An algorithm is considered the best if it can minimize the error in previous iterations [34].

The main idea of AdaBoostRegressor is to combine weak classifiers obtained in the iterative process during which a new model learns at each step taking into account the data on the errors of the previous models [35].

The three aforementioned methods are based on the gradient boosting. Similarly, to any other boosting algorithm, it consistently builds basic models in a way that each subsequent one improves the quality of the whole ensemble. Gradient boosted decision trees algorithm builds a model as a sum of trees (4):

$$f(x) = h_0 + \vartheta \sum_{j=1}^{M} h_j(x), \qquad (4)$$

where h_0 is a constant model (initial initial guess), $\vartheta \in (0, 1)$ is the parameter which regulates the learning speed and the influence of individual trees on the whole model, and $h_j(x)$ is regression trees.

The results of the aforementioned models are presented in the Table I.

TABLE I. FORECAST RESULTS

Model	Chelyabinsk	Khanty-Mansiysk
	Coefficient of determination (R^2)	
LinearRegression	0.8353	0.4460
CatBoostRegressor	0.9004	0.8605
XGBRegressor	0.8873	0.8535
AdaBoostRegressor	0.8820	0.8394

In both cases, the CatBoostRegressor model showed the best results, while the prediction rates of XGBRegressor and AdaBoostRegressor are slightly lower. Notably, the linear regression model showed a better result for Chelyabinsk than for Khanty-Mansiysk, however, compared to the other models, its forecast is significantly worse. This can be ascribed to the DataFrame size. The quality of prediction for Khanty-Mansiysk can be improved by increasing the data size.

The root-mean-square error (RMSE) for Chelyabinsk is 287,785 rubles and for Khanty-Mansiysk is 515,697 rubles. It means that the difference between the predicted and actual prices cannot exceed these amounts. The relationship of the predicted values to the actual values for Chelyabinsk and Khanty-Mansiysk is illustrated in the graphs (Fig. 5 and Fig. 6).

The diagonal line in the graph signifies where all the points would lie if there was a perfect match between the actual and the predicted values.

V. DETERMINING KEY CHARACTERISTICS

DataFrame includes 44 characteristics, and it is necessary to evaluate their importance for the price. CatBoostRegressor showed the least error rate in the forecast, therefore, this model was used as a basis for determining the importance of characteristics (Fig. 7 and Fig. 8).

Feature importance ranking for Chelyabinsk illustrated that the total living space of an apartment, Tsentralny district, number of floors, commissioning year, load-bearing panel walls and other features affected the price most significantly.

Feature importance ranking for Khanty-Mansiysk showed that the total living space of an apartment, living space, floor number, number of floors, cellar space and other features affected the price most significantly.

In Chelyabinsk, the total living space of an apartment (51.5) and the district (18.5) have the biggest influence on the price. As for Khanty-Mansiysk, the importance of the total living space is bigger (59) and the district is not included in the top ten important characteristics as only two districts can

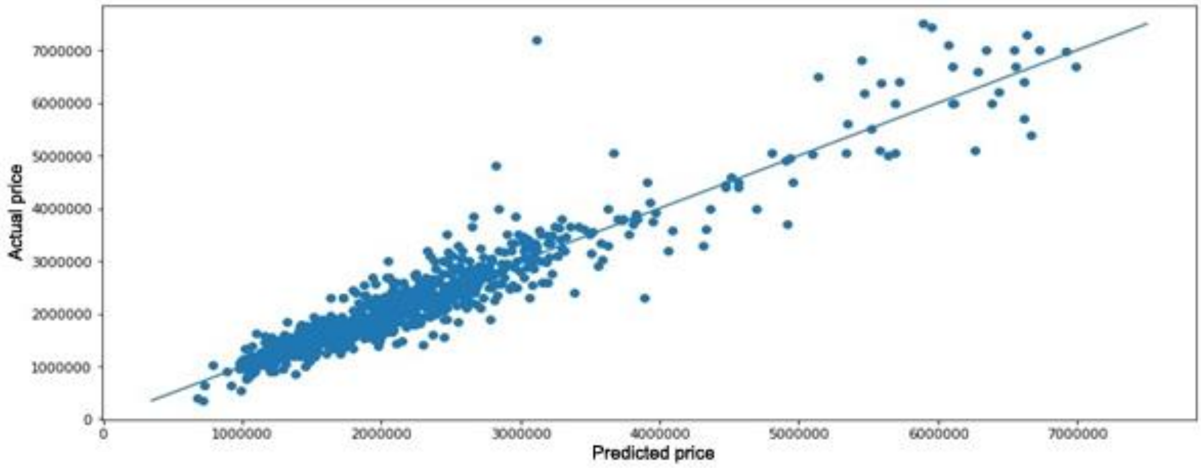


Fig. 5. Relationships between values for Chelyabinsk.

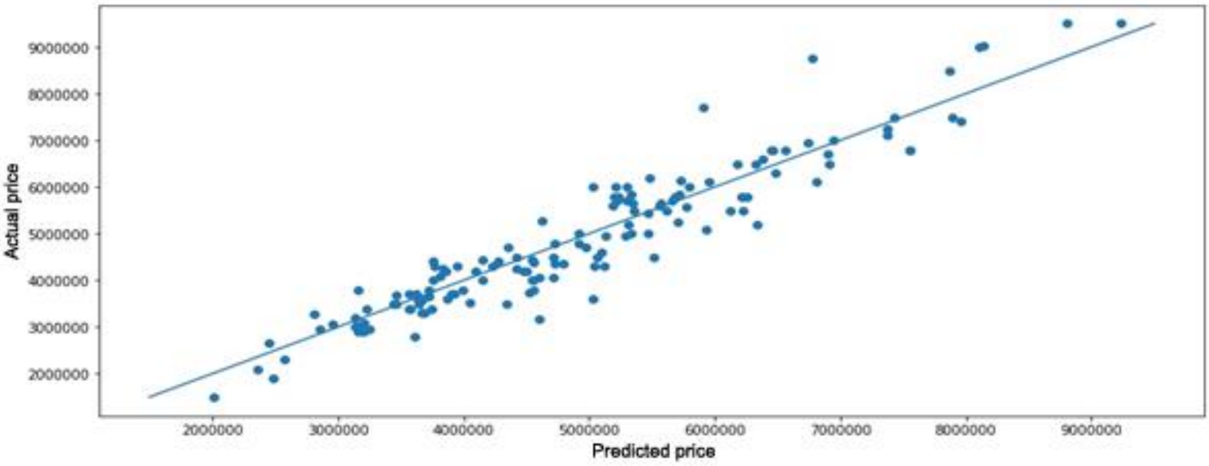
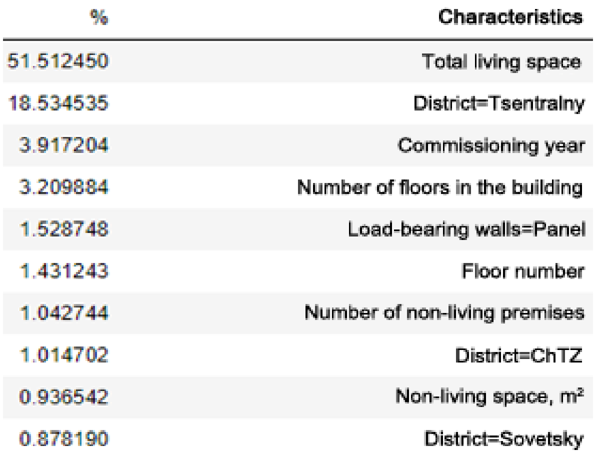


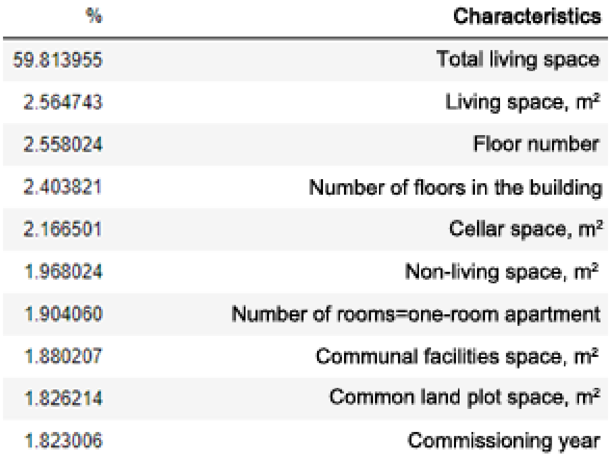
Fig. 6. Relationships between values for Khanty-Mansiysk.

be selected and the city is relatively small, compared to Chelyabinsk.



%	Characteristics
51.512450	Total living space
18.534535	District=Tsentralny
3.917204	Commissioning year
3.209884	Number of floors in the building
1.528748	Load-bearing walls=Panel
1.431243	Floor number
1.042744	Number of non-living premises
1.014702	District=ChTZ
0.936542	Non-living space, m²
0.878190	District=Sovetsky

Fig. 7. Importance of characteristics for Chelyabinsk.



%	Characteristics
59.813955	Total living space
2.564743	Living space, m²
2.558024	Floor number
2.403821	Number of floors in the building
2.166501	Cellar space, m²
1.968024	Non-living space, m²
1.904060	Number of rooms=one-room apartment
1.880207	Communal facilities space, m²
1.826214	Common land plot space, m²
1.823006	Commissioning year

Fig. 8. Importance of characteristics for Khanty-Mansiysk.

VI. Conclusion

From the analysis, it was concluded that the average price of apartments in Chelyabinsk was significantly lower than in Khanty-Mansiysk. In spite of that, the housing affordability index in both cities is equal to 2. Moreover, the residential stock for sale in Khanty-Mansiysk is newer. Price forecasting models used in the study demonstrated that CatBoostRegressor created by the technology company Yandex had the best rate. Other models, with the exception of the linear regression, showed almost the same result which was a few percent less. Linear regression model illustrated the worst result, moreover, its prediction quality level for Khanty-Mansiysk was less than 50%, probably, due to the substantial amount of data. Feature importance ranking permitted to identify that in Chelyabinsk, the total living space of apartments and their location in Tsentralny district played an important role in predicting the price, while the main characteristic for Khanty-Mansiysk was the total living space.

Future work should concentrate on increasing the amount of data and testing other machine learning methods to improve the forecast. Presumably, using other techniques or increasing the amount of data on residential properties would allow to improve the quality of the price forecast to 95% or higher, given that the levels of 90% and 86% for Chelyabinsk and Khanty-Mansiysk respectively are not high enough.

References

[1] "How a developer should work with escrow accounts" (in Russian). Unified Information System for Housing Construction. https://наш.дом.рф/press/article/2019/04/kak-zastrojshchiku-rabotat-so-schetami-eskrou1 (accessed Jan. 12, 2020).

[2] E. Miroshkina, "Standard VAT to be increased to 20%. What it actually means" (in Russian). Journal.Tinkoff.ru. https://journal.tinkoff.ru/fake-news/nds-20-18/ (accessed Jan. 14, 2020).

[3] "Housing loans in rubles granted to resident individuals" (in Russian). CBR.ru. https://cbr.ru/statistics/table/?tableId=4-1 (accessed Jan. 18, 2020).

[4] "Mortgaging in figures. Mortgage loan granting statistics" (in Russian). Rusipoteka.ru. http://rusipoteka.ru/ipoteka_v_rossii/ipoteka_statitiska/ (accessed Jan. 14, 2020).

[5] E. Komissarova, "Assesment of the Russian housing market current state: Construction volume and price forecasts" (in Russian). Rusipoteka.ru. http://rusipoteka.ru/files/events/2019/2609/komissarova-ocenka-sostoyaniya-rinka-giliya-rf.pdf (accessed Jan. 15, 2020).

[6] K. Farrelly, and S. Stevenson, "The risk and return of private equity real estate funds," *Global Finance Journal*, vol. 42, 2019.

[7] L. d'Acci, "Quality of urban area, distance from city centre, and housing value. Case study on real estate values in Turin," Cities, vol. 91, pp. 71–92, 2019.

[8] M. Renigier-Biłozor, A. Janowski, and M. d'Amato, "Automated Valuation Model based on fuzzy and rough set theory for real estate market with insufficient source data," *Land Use Policy*, vol. 87, 2019.

[9] H. Selim, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," *Expert systems with Applications*, vol. 36, no. 2, pp. 2843–2852, 2009.

[10] A. Khalafallah, "Neural network based model for predicting housing market performance," *Tsinghua Science and Technology*, vol. 13, no. S1, pp. 325–328, 2008.

[11] Yeh I., C. Yeh, and T. K.Hsu, "Building real estate valuation models with comparative approach through case-based reasoning," *Applied Soft Computing*, vol. 65. pp. 260–271, 2018.

[12] B. Park, and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol. 42. pp. 2928–2934, 2015.

[13] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3383–3386, 2011.

[14] J. J. Ahn, H. W. Byun, K. J. Oh, and T. Y. Kim, "Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8369–8379, 2012.

[15] E. A. Antipov, and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1772–1778, 2012.

[16] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO*,*" *Optik*, vol. 125, pp. 1439–1443, 2014.

[17] M. Cupal, "The Comparative Approach theory for real estate valuation," Procedia-Social and Behavioral Sciences, vol. 109, pp. 19–23, 2014.

[18] "Information about houses in Russia" (in Russian). Dom.MinGKH.ru. http://dom.mingkh.ru/ (accessed Jan. 10, 2020).

[19] L. Richardson, "Beautiful Soup Documentation." Leonard Richardson's personal website Crummy.com. https://www.crummy.com/software/BeautifulSoup/bs3/documentation.html (accessed Jan. 10, 2020).

[20] K. Reitz, "Requests: HTTP for humans." PyPi.org. https://pypi.org/project/requests/ (accessed Jan. 10, 2020).

[21] "CSV file reading and writing." Python 3.8.2 documentation. https://docs.python.org/3/library/csv.html (accessed Jan. 10, 2020).

[22] M. Driscoll, “Python sleep(): How to add time delays to your code.” RealPython.com. https://realpython.com/python-sleep/ (accessed Jan. 10, 2020).

[23] “DomClick – searching for and getting a mortgage on an apartment” (in Russian). DomClick.com. https://domclick.ru/ (accessed Jan. 10, 2020).

[24] “Free Proxy List – Just checked proxy list.” Free-Proxy-List.net https://free-proxy-list.net/ (accessed Jan. 10, 2020).

[25] “Nominatim Documentation.” Nominatim.org. http://nominatim.org/release-docs/develop/ (accessed Jan. 10, 2020).

[26] “Russians still choose residential property based on the price and location” (in Russian). NAFI.ru. https://nafi.ru/analytics/rossiyane-po-prezhnemu-vybirayut-zhile-po-tsene-i-raspolozheniyu/ (accessed Jan 28, 2020).

[27] “Developer’s story is a key factor when choosing an apartment in a new-built” (in Russian). NAFI.ru. https://nafi.ru/analytics/istoriya-zastroyshchika-vazhnyy-faktor-vybora-kvartiry-v-novostroyke-en-history-of-construction-deve/ (accessed Jan 28, 2020).

[28] “Housing Affordability Index – Methodology.” NAR.realtor. https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index/methodology (accessed Jan 29, 2020).

[29] “Average price of 1 square meter of the total living space of apartments in the market” (in Russian). Federal Statistics. https://www.fedstat.ru/indicator/31452 (accessed Jan 27, 2020).

[30] “Job market, employment, and wages” (in Russian). Federal State Statistics Service. http://old.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/wages/ (accessed Jan 27, 2020).

[31] A. C. Müller, and S. Guido, Introduction to Machine Learning with Python. Sebastopol, CA, USA: O’Reilly Media, 2017, p. 392.

[32] “CatBoostRegressor – CatBoost. Documentation.” CatBoost.ai. https://catboost.ai/docs/concepts/python-reference_catboostregressor.html (accessed Jan 17, 2020).

[33] “Developers on CatBoost” (in Russian). Yandex.ru. https://yandex.ru/dev/catboost/ (accessed Jan 18, 2020).

[34] A. Khapkin, “Writing XGBoost from scratch part 2: gradient boosting” (in Russian). Mail.ru Group’s blog on Habr. https://habr.com/ru/company/mailru/blog/438562/ (accessed Jan 16, 2020).

[35] “Boosting, AdaBoost” (in Russian). ITMO University. https://neerc.ifmo.ru/wiki/index.php?title=Бустинг,_AdaBoost (accessed Jan 18, 2020).