Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 10, No. 2, September 2021, pp. 13-18

Effective use of Machine Learning to Solve Socio-economic Problems

[1*]Sapakova Saya, [2] Atudinov Dias., [2]Abdikadirov Arnur, [2]Yilibule Yelidana., [2]Ybyrakhym Nuray

[1]International IT University, Almaty, Kazakhstan
[2]Kazakh National University named after Al-Farabi, Almaty, Kazakhstan

*Corresponding Author: s.sapakova@iitu.edu.kz

Article Info

Article history:

Article received on 28 12 2021

Received in revised form 29 12 2021

Keywords:

real estate market, cost of apartments, real estate object, regional real estate market, Almaty, machine learning, algorithms.

ABSTRACT: To study of the dependence of the offer price of a residential real estate object in Almaty on a number of factors (apartment area; number of rooms; material from which the house is built; floor on which the studied object is located; kitchen area; type of bathroom; presence of a balcony; distance from the city center; availability of an elevator; condition of the apartment in terms of the need for repairs; age of the house; belonging to the primary or secondary real estate market). The obtained machine learning models can be used to make forecasts of the cost of residential real estate, which can be reflected in the development of programs for the socio-economic development of municipalities and the region as a whole, will be in demand by investors, other counterparties in the residential real estate market and individuals. when assessing the acquired social and domestic benefits. The authors see the continuation of research based on the accumulation of an array of information in expanding the possibilities of modeling the cost of apartments, depending not only on the characteristics of apartments, but also on characteristics that reflect the influence of external factors on the regional market of residential real estate.

1. INTRODUCTION

The relevance of this study is associated with the special social significance of the real estate market and the high influence of this market on the components of other segments of the market economy. It is generally accepted that the achieved level of market transformations in the housing and communal sector and the sphere of real estate turnover as a whole determines the degree of development of market mechanisms in the national economy.

Distinctive features of the Kazakhstan residential real estate market from similar markets in other countries, where lending systems for real estate transactions are developed, are:

- excess of supply over effective demand;

- dissatisfaction with the housing conditions of the majority of the country's population; weak development or virtually no competition in the primary housing markets; significant disparities in the development of market mechanisms and economic and legal regulation of the housing market between individual regions.

The latter circumstance dictates the need for a detailed study of the patterns of formation and features of the functioning of regional real estate markets, in which market mechanisms of interaction between market entities are most manifested. These are, first of all, the main strategically important cities: Nur-Sultan, Almaty, Shymkent.

In the works of foreign scientists, much attention is paid to the study of price fluctuations in the real estate market and the reasons that cause them. The main methods of analysis and forecasting of the situation on the foreign real estate market are statistical methods, which show themselves well in stable economic conditions, since the real estate market has a fairly large inertia. For countries experiencing rapid economic growth or crisis, statistical models cannot be used to obtain adequate results, as indicated in the works of C. Brooks, S. Tsolacos [1]. Nevertheless, these scientists widely use traditional modeling and forecasting tools (regression analysis, parametric models, linear models, various types of moving averages, etc.). S.J. Maisel, J.B. Burnham [2], J.S., A.M. Polinsky, D.T. Ellwood [3], F. De Leeuw [10] describe the features of using the concept of elasticity in the analysis of the real estate market and draw conclusions about the fundamentally different nature of elasticity in the short and long term.

J.W. Forrester describes the methods of simulation modeling. Korean scientists develop his ideas D. Cho, S. Ma [9], S.-J. Hwang, M.-S. Park, H.-S. Lee, Y.-S. Yoon [10], their work on modeling the national real estate market is distinguished by deep study and a large number of factors taken into account. M. Eskinasi systematized the main results of the analysis of system dynamics methods in the study of the real estate market. Consider some of the methods proposed by researchers to model the cost of apartments.

Sidorenko O.A. in article [12] propose a multilevel (hierarchical) approach to modeling the cost of residential real estate at the regional level: the price of a real estate object is formed simultaneously under the influence of the characteristics of the object itself (microlevel - apartment area, balcony, floor) and characteristics the socio-economic state of the administrative territorial unit in which it is located (meso-level - the level of wages, unemployment, distance to the regional center). The authors of the article use a small number of factors to assess but interesting is the introduction to the model of the meso-level indicators.

The authors of the article use a small number of factors to assess, but it is interesting to introduce meso-level indicators into the model. The study was carried out on the basis of data from Almaty and Almaty oblast. There are also concepts that imply a greater number of factors, divided into three hierarchical levels: country (social, economic, political factors), local (features of the location of the object, terms of sale, temporary) and individual visual (architectural and construction and financial and operational features). Unfortunately, due to the complexity of determining the listed factors, the application of this method is limited.

In the works of A.Yu. Mints [11], the use of dynamic simulation modeling for the analysis and forecasting of pricing in the Ukrainian residential real estate market is considered. The author of the work analyzes the factors acting in the secondary real estate market, builds models of causal relationships between demand, supply and price, as well as between factors that form the financial capabilities of buyers of the real estate market. Then, on their basis, dynamic simulation models of pricing are built and experiments are carried out confirming that the most significant factor in pricing in the real estate market is external sources of financing (bank credit). V.L. Yasnitsky proposes the use of neural network modeling for the mass appraisal of residential real estate [21]. He developed a computer program, in which is based on a neural network trained on the results of free information resources containing information about the real estate offer price. After excluding outliers associated with information uncertainty, the researcher published the average relative error of neural network prediction results, which is at the level 1.03%. Using a neural network model, he estimates the significance of the input parameters, highlighting the most important of them, and predicts the market value of residential real estate in the city of Perm.

To determine the cost of real estate in housing construction, costly, profitable, comparative approaches are most often used. Since the interests of various economic subjects of the national economy are manifested in the housing sector: business, households, the state, it is with their consideration that the final cost of residential real estate is formed. Note that at present, the real level of wages of the general population does not allow speaking about the affordability of housing [2, 3, 5]. I.V. Burova, M.V. Panichkin [5-6] check the adequacy of the application of econometric analysis methods to assess the cost of real estate in Rostov-on-Don and build their own cost models. Researchers convincingly prove that the constructed models are effective separately for groups by the number of rooms, reflect the market value of a residential property in conditions of uncertainty, and come to the conclusion that the use of econometric analysis methods is justified for calculating the market value of real estate, if they do not have significant distortions in a crisis. Thus, a brief review of the literature over the past 10 years on the problem under study has shown that researchers use different methods and models with varying degrees of success to build models and predict prices in the real estate market. It should be noted that in the scientific literature to study the real estate market, either traditional econometric methods and models - factorial, statistical, etc., are used, or the synthetic method, in which the cost of real estate is considered as an integral assessment of the cost of various factors [9-11].

2. PROBLEM STATEMENT

Each property is described by a set of characteristics, the values of which affect its market value. The task is to use data on various objects, including, on the one hand, the values of the characteristics of each object, and on the other, the value of its market price, to propose a rule with which it is possible to predict the most probable market price of a new th object according to its characteristics. Note that we are talking about all the characteristics that affect its market value, including the technical parameters of the object itself, characteristics of its location and factors of the market environment in which the objects are sold.

The objectives of this work are to analyze the fundamental possibility of determining the market value of real estate objects based on AI technologies, to evaluate the assessment results according to generally accepted accuracy criteria by different methods (algorithms) and select the most effective of them, to study the effect of parameter of the selected algorithm on the accuracy of the result. The research carried out in the article is based on the market data of the city of Almaty. At the same time, the general conclusions obtained as a result of the experiments can be attributed to other regions of the Republic of Kazakhstan. To carry out the experiments within the framework of this article, we used data from real ads from the category "sale of secondary housing" from the site krisha.kz using data parsing.

3. MATERIALS AND METHODS

The purpose of this work is to study the dependence of the cost of apartments in Almaty on a number of factors (apartment area; number of rooms; material from which the house is built; floor on which the object under study is located; kitchen area; type of bathroom; availability balcony; remoteness from the city center; availability of an elevator; condition of the apartment in terms of the need for renovation; age of the house; belonging to the primary or secondary real estate market).

The resulting sample was divided into two parts: training (70%) and test (30%). The training sample was used to train the models, and the test sample was used to determine the quality of their prediction. Let us determine the prices of offers posted on the sites for the sale of residential real estate as y, and the predicted values of the value of residential real estate as ŷ. To assess the effectiveness of the procedure used, the following accuracy characteristics (metrics) are calculated for each result:

1) the coefficient of determination (R2) reflects the share of the explained variance of the model.

The closer the coefficient of determination is to 1, the stronger the agreement of the model with the data.

2) the average absolute error (mean ab solute percentage error - MAPE) shows how many percent the model is wrong on average.

3) median absolute error (medi an absolute percentage error - MedAPE) reflects the mean value among all ordered values of percentage errors.

The specified characteristics were calculated separately for the sample used for training and for the selected test sample. We draw attention to the fact that the test sample did not participate in the training in any way; therefore, the accuracy characteristics calculated by comparing the predicted values of the market value and the actual prices of bids for the same objects reflect the real accuracy of the prediction. The estimation accuracy depends to a large extent on the set of features (price-forming parameters), with the help of which the object of assessment is identified, therefore, an important part of the formation of initial data is the determination of the composition of features for describing each object.

It should be noted that the completeness of the description of the object is limited to information about the objects that are contained in the advertisements for sale posted on the corresponding resources.

A meaningful analysis of the housing market made it possible to single out the following essential characteristics of an object that determine its market value:

1) numeric variables:
- year of construction;
- number of storeys;
- the total area of the apartment;
- kitchen area;

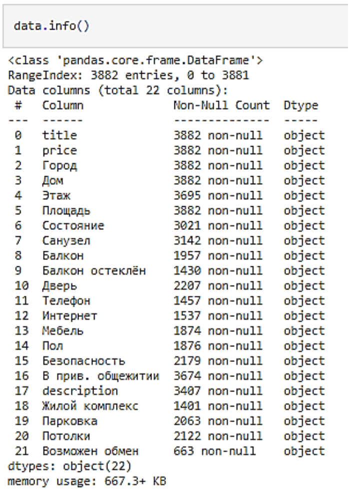
2) categorical variables:
- district;
- number of rooms;
- wall material;
- placement floor;
- safety
- territorial zone.

Note that the specific set of parameters may differ depending on the site from which the pricing data was obtained.

Usually, most of the characteristics are contained in advertisements posted on various sites for the sale and rent of different types of real estate. However, the advertisements do not always indicate the important features of the objects for sale. These signs are most fully presented on the site krisha.kz. But even on this site there is significant information about the location and condition of the apartment not fully presented.

As for the location, within the framework of this study, it was customary to characterize it by the price zone in which the object is located. The zoning required for this can be accomplished by clustering according to various criteria, for example, using the coordinates of objects.

Another problem that should be addressed at the stage of data preparation is related to the presence of numerous "obstacles" in the ads (outliers, false ads with overpriced and underpriced prices, with inadequate and contradictory characteristics of the object, etc.). Due with this, the sample formed from the market data was properly prepared for the research - outliers, ads with inadequate data, with very low or very high prices were removed. When processing numeric characters, values less than 2 percentile and more than 98 percentile were discarded. All values of numeric features were reduced to a single type, for example, to integer or real. Categorical signs also require standardization - all letters are reduced to lowercase so that "Brick" walls do not differ from brick walls ", extra spaces and punctuation marks are removed, spelling mistakes made by the seller are corrected. Some characteristics require more specific transformations, for example, the values of the attribute "location floor" of a real estate object were transformed into three types: first floor, middle and last.



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3882 entries, 0 to 3881
Data columns (total 22 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   title              3882 non-null   object
 1   price              3882 non-null   object
 2   Город              3882 non-null   object
 3   Дом                3882 non-null   object
 4   Этаж               3695 non-null   object
 5   Площадь            3882 non-null   object
 6   Состояние          3021 non-null   object
 7   Санузел            3142 non-null   object
 8   Балкон             1957 non-null   object
 9   Балкон остеклён    1430 non-null   object
 10  Дверь              2207 non-null   object
 11  Телефон            1457 non-null   object
 12  Интернет           1537 non-null   object
 13  Мебель             1874 non-null   object
 14  Пол                1876 non-null   object
 15  Безопасность       2179 non-null   object
 16  В прив. общежитии  3674 non-null   object
 17  description        3407 non-null   object
 18  Жилой комплекс     1401 non-null   object
 19  Парковка           2063 non-null   object
 20  Потолки            2122 non-null   object
 21  Возможен обмен     663 non-null    object
dtypes: object(22)
memory usage: 667.3+ KB
```

Figure 1: Validation accuracy

In this paper, we consider 22 factors that describe real estate in the city of Almaty in Bostandyk district.

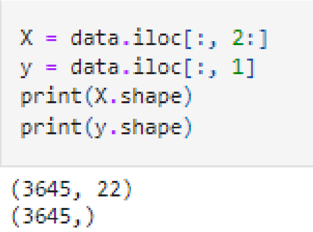
3.1. Data preprossesing

One of the most challenging tasks faced by the evaluator is structuring the information derived from the ad text. The fact is that the most significant information about the value of the object of appraisal is in the content of the ad, which is usually written in an arbitrary form in accordance with the personal preferences of the author of the ad (seller, etc.). With proper processing, this description can be used to form the necessary features and thus improve the accuracy of model estimation. During the course of this study, various relevant features were identified. This made it possible to present the description of the object in a structured form and to provide further processing this information. The article will give only one example of using a detailed text description to obtain a new feature.

An important characteristic influencing the cost of an apartment is its condition. For the problem in this example, the following target classes were defined that characterize the condition of the apartment:

- without repair;
- typical repair;
- renovation;
- author's project.

The website from which the information for research was obtained did not have such information, but a detailed description of the object offered for sale was offered. To form the training sample, we used the sites on which the state was indicated in accordance with this classification. Having such a database, the problem of identifying the state of an object based on a text description can be reduced to the problem of classification into several classes. The test sample in our case is a database of ads for residential real estate in Almaty with a detailed description.

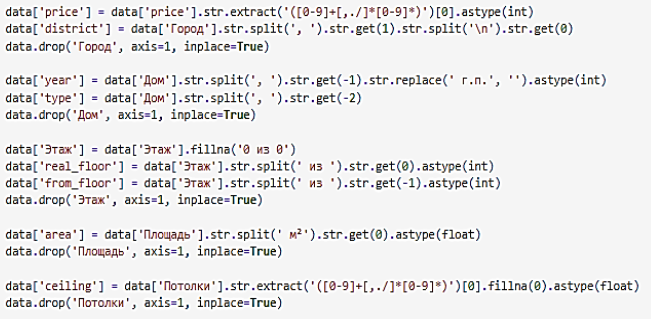


```
X = data.iloc[:, 2:]
y = data.iloc[:, 1]
print(X.shape)
print(y.shape)
```

```
(3645, 22)
(3645,)
```

Figure 2: Data size

At the stage of data preprocessing, we cleaned irrelevant data, and also split the composite values of characteristics into separate words for convenience and structuring of information. As you can see from the Figure 3.



```
data['price'] = data['price'].str.extract('([0-9]+[,./]*[0-9]*)')[0].astype(int)
data['district'] = data['Город'].str.split(', ').str.get(1).str.split('\n').str.get(0)
data.drop('Город', axis=1, inplace=True)

data['year'] = data['Дом'].str.split(', ').str.get(-1).str.replace(' г.п.', '').astype(int)
data['type'] = data['Дом'].str.split(', ').str.get(-2)
data.drop('Дом', axis=1, inplace=True)

data['Этаж'] = data['Этаж'].fillna('0 из 0')
data['real_floor'] = data['Этаж'].str.split(' из ').str.get(0).astype(int)
data['from_floor'] = data['Этаж'].str.split(' из ').str.get(-1).astype(int)
data.drop('Этаж', axis=1, inplace=True)

data['area'] = data['Площадь'].str.split(' м²').str.get(0).astype(float)
data.drop('Площадь', axis=1, inplace=True)

data['ceiling'] = data['Потолки'].str.extract('([0-9]+[,./]*[0-9]*)')[0].fillna(0).astype(float)
data.drop('Потолки', axis=1, inplace=True)
```

Figure 3: Fragment of the data preprossesing

Training data size: 2916, and test data: 729 for this dataset.

3.2. Applying DecisionTreeRegressor algorithm

There are many machine learning methods that in principle can be used to provide a process for determining the market value of a property. This section presents the results of studies of various machine learning methods, including linear regression, which is widely used in evaluative practice, classical machine learning algorithms (random forest, gradient boosting), and more modern models (xgboost, catboost) ... In addition, the results of the assessment based on the use of a neural network are considered. Thus, the following methods are considered in the work:

- Linear Regression;
- Decision Tree.

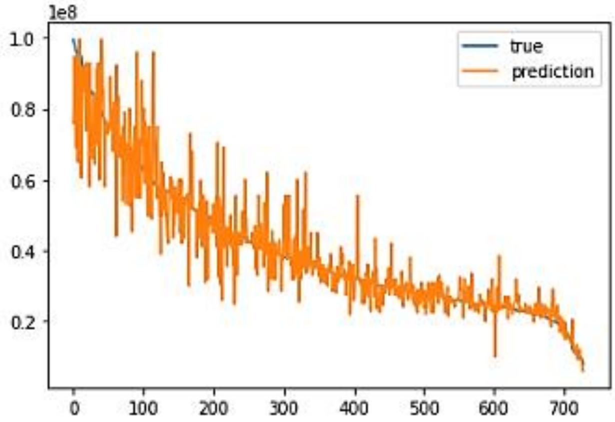


Figure 4: DecisionTreeRegressor plot

Analyzing the correlation matrix, we selected 6 main factors and applied the above algorithm again.

Left behind the following factors: 'year', 'from_floor', 'area', ‘Bathroom ',' Residential complex '.

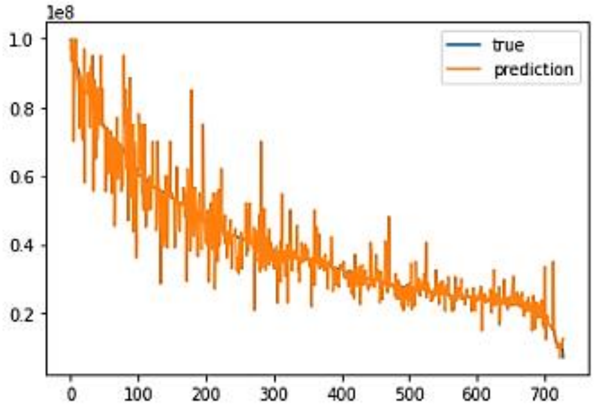


Figure 5: DecisionTreeRegressor plot for six features

In the first case, when 20 factors were used, we received the following estimates: train accuracy: 0.9999830490792269 and test accuracy: 0.8602945585231344.

In the second case, when 6 main factors were used, we received the following estimates: train accuracy: 0.9977983354961577 and test accuracy: 0.8801565538608795. So 20.713305898491083 is exactly the same percentage and the rest is 51167321553660.29 mse.

As already noted, for an adequate assessment, a sufficiently complete description of the object is required. Identification degree an asset as an object of valuation is determined by the completeness of its description using price-generating parameters. Moreover, an important role is played not only by the parameters, but also by the degree of influence of each factor taken into account in the assessment of the factor on the market value of the object. Our analysis of the data made it possible to determine the significance (weight) of each of the features used in terms of its impact on the value of the object.

Table 1. The significance of features from the point of view of its influence on the value of the object

Features	Significance
District	23.73
Built in	16.85
Territorial zone	12.86
Number of storeys	10.52
Total area	8.01
Repair	7.68
Wall material	6.48
Kitchen area	5.78
Number of rooms	5.03
Accommodation floor	3.0

The first computational experiments were carried out taking into account all factors.

4. CONCLUSION

Within the framework of this article, the problem of determining the value of residential real estate by its characteristics was considered. To solve this problem, machine learning and data mining method was used. During the research, we collected and the data posted on the sites for the sale of residential real estate were analyzed, several machine learning algorithms were trained, several additional features were formed that improved the quality of the models, the best method was chosen, based on which the dependences of quality metrics on various features were analyzed. The studies carried out made it possible to obtain the accuracy 0.88 on a test sample using the DecisionTreeRegressor algorithm with main features.

The research results obtained in the framework of this work confirm the effective application of machine learning for solving the problem of determining the value of residential real estate objects. The developed algorithms can be implemented in the work of appraisal companies to improve the quality of appraisal reports. In addition, the methods used can be applied to determine the prices of commercial real estate, for mass estimates of the cadastral value. For future research, to improve the accuracy of the estimate, you can work in more detail on the definition of the type of repair, using the detailed description from the announcements: use a larger training set and

mark up the data to obtain reliable information. For the same purpose, you can try to apply a convolutional neural network, which will determine the type of apartment renovation based on photographs from the ad.

REFERENCES

[1] Elaine M. Worzala, Margarita Lenk, Ana Silva. An Exploration of Neural Networks and Its Application to Real Estate Valuation // Journal of Real Estate Research ; American Real Estate Society, vol. 10(2). pp. 185–202, 1995.

[2] Visit Limsombunchai (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, American Journal of Applied Sciences. 1(3). pp. 193–201.

[3] G. Bernard, J. S. Pathmanathan, R. Lannes, P. Lopez, and E. Bapteste, “Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery,” Genome Biol. Evol., vol. 10, no. 3, pp. 707–715, Mar. 2018, doi: 10.1093/gbe/evy031.

[4] GeoPhy : [site]. URL: https://geophy. com/

[5] Yasnitskiy V.L. Neural network modeling in the problem of mass appraisal of residential real estate in the city of Perm // Fundamental research. 2015. No. 10-3. pp. 650–653. URL: http://www.fundamental-research.ru/ru/article/view?id=39274

[6] Surkov F. A., Petkova N. V., Sukhovsky S. F. Neural network methods of data analysis in real estate appraisal "// Izvestia universities. North Caucasian region. Technical sciences. 2016. No. 3. pp. 38-45.

[7] Aref'eva E.A., Kostyaev D.S. Using neural networks to assess the market value of real estate, Bulletin of the Tula State University. Technical science. 2017. Issue. 10, pp. 177–184.

[8] HouseCanary - Residential real estate valuations: [website]. URL: https://www.house canary.com/.

[9] Cho D., Ma S. Dynamic Relationship between Housing Value and Interest Rates in the Korean Housing Market // The Journal of Real Estate Finance and Economics. 2006. Vol. 32, № 2. P. 169–184. DOI: 10.1007/s11146-006-6013-6.

[10] Hwang S.-J., Park M.-S., Lee H.-S., Yoon Y.-S. Analysis of the Korean Real Estate Market and Boosting Policies Focusing on Mortgage Loans: Using System Dynamics // Korean Journal of Construction Engineering and Management. 2010. Vol. 11, № 1. P. 101–112. DOI: 10.6106/kjcem.2010.11.1.101

[11] Mints A.U. Modeling of the pricing process in the housing market by the methods of system dynamics // Technology audit and production reserves. 2016. T. 5, № 4 (31). pp. 39–45..

[12] Sidorenko O.A. The main directions of economic and mathematical modeling of the real estate market. Statistika i ekonomika [Statistics and Economics], 2013, no. 3, pp. 153–158. (In Russian)