ESTIMATING MARKET VALUE OF APARTMENTS USING THE k-NEAREST NEIGHBORS ALGORITHM

Marko Dragojević [1]
Nikola Stančić [2]

UDK: 332,622
DOI: 10.14415/konferencijaGFS2019.098

Summary: *The market value of apartments is, as the name itself suggests, defined by the sellers and the buyers through supply and demand – elements that collectively make up the market. Observing a large number of factors affecting the price of real estate is not an easy job. Price formation depends on both the characteristics of the apartment and the buyer's value-system. The basic question that a rational customer asks himself is "why would I pay a larger sum of money for the same or practically same thing than what someone else paid for it just recently?". This fact leads to the conclusion that it is necessary to know the characteristics and prices of the real estates traded in the near past and in the close surrounding. A comparative way of customer's thinking is the basic principle for defining one such model. This is a necessary but not sufficient condition. Models based on the machine learning algorithms (among them k-Nearest Neighbors algorithm) require having a larger amount of data, so that the made conclusions can be reliable, accurate, and precise.*

Keywords: *estimating market value, market of apartments, data mining, k-Nearest Neighbors algorithm*

1. INTRODUCTION

Market value is the most probable amount for a transaction between an average buyer and an average seller who are in an arm's-length transaction after proper marketing ("exposure to the market"), where both sides are well informed and behave reasonably and prudently, with the seller and the buyer not being forced to sell or buy [1].
Therefore, estimating the market value of a real estate is a very complex business. It is necessary to have a broader image about state and trends in a particular market. Questions that arise are what items are in high demand, what is currently selling well, necessarily taking into account the whole range of prerequisites, conditions of restriction, etc. Technological characteristics are only one aspect to be considered, the most important ones are usually of other nature. The assumption is that the characteristic most affecting the value of the apartment is its location, followed by the micro-location and the apartment's orientation. This means that, if all the essential features of the

[1] Marko Dragojević, mast. inž. građ., University of Belgrade, Faculty of Civil Engineering Beograd, Bulevar Kralja Aleksandra 73, Belgrade, Serbia, tel: +381 63 84 82 640, e – mail: markomionica@gmail.com
[2] Nikola Stančić, mast. inž. geod., University of Belgrade, Faculty of Civil Engineering Beograd, Bulevar Kralja Aleksandra 73, Belgrade, Serbia, tel: +381 62 746 453, e – mail: nstancic@grf.bg.ac.rs

apartments on one hand, and the desires and capabilities of the buyers on the other hand are known, it would be possible to assess the value of a particular apartment with great precision. In situations where the available data has the knowledge of certain significant facts covered, an excellent solution for predicting the unknown value of some apartment would be a method that includes making conclusions based on the data or the application of some of the methods of machine learning.

Data mining, also known as data analysis, can be defined as the process of analyzing large information repositories and of discovering implicit, but potentially useful information [2]. It owes its origins to a process known as the knowledge discovery in databases (KDD) [3], in which it originally represented one of the steps. The main characteristic of data mining is the capability to uncover seemingly hidden relationships and to reveal unknown patterns and trends by 'digging' into large amounts of data [4]. These detected patterns can represent some new knowledge that, as such, can be used to predict the future behavior of a certain phenomenon. Practically, data mining usually involves analyzing the vast amount of historical data which was previously considered unusable and was ignored.

The last decades have brought significant advances in techniques relating to the collection and storage of various data. This progress has led to a sort of a flooding with data. However, this amount of data is not being accompanied by the proportional acquisition of new knowledge. With the development of the technology and the increase of the processing power of computers, it's become easier to process larger sets of data in a more effective way. For this reason, data mining today represents an important method for modeling something *unknown* on the basis of something *known* that, at first glance, seems like non-informative data.

With all of this in mind, the aim of this paper is to create a model that will predict the price or class of the apartment based on the previous experience and the known features of other apartments. Algorithms belonging to machine learning will be applied for these purposes. In particular, a practical example is done using the k-Nearest Neighbors approach.

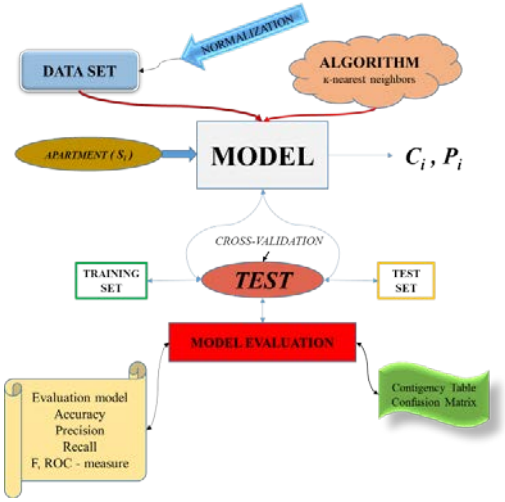

Figure 1. Graphic representation of the model

2. MATHEMATICAL BASIS OF THE k-NEAREST NEIGHBORS ALGORITHM

k-Nearest Neighbors algorithm (abbr. KNN) is one of the most used machine learning algorithms. The idea behind this algorithm is to mimic a comparative way of thinking. Thus, it is very similar to the most used approach for assessing the value of real estate – comparative method. The algorithm itself is based on storing objects in the m-dimensional space, where the number *m* represents the number of different characteristics (attributes). Since the object has certain attribute values, it occupies a specific point in this space.

If a large number of objects is collected and placed in m-dimensional space, it can be noticed that similar objects are grouping together. These groups of objects will most likely represent separate classes. Given the position of a new object whose class is not known, the algorithm will be able to determine what class it most likely belongs to. The criteria for predicting its most probable class are the classes for *k* other objects from the immediate environment. The basic idea of this algorithm is shown in *Figure 2*.

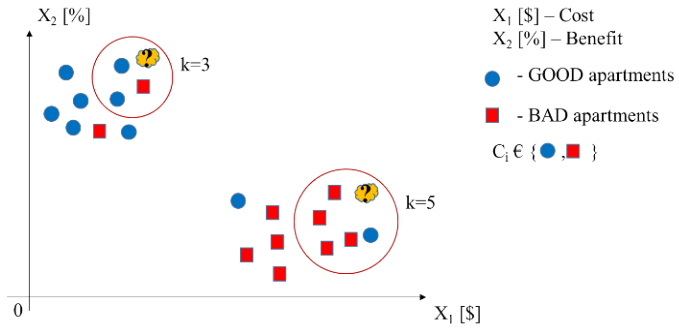

Figure 2. Example of the k-Nearest Neighbors classifier

The number of nearest neighbors must be an odd number (usually 1, 3, 5 or 7), in order to be able to choose the winning class. The closer neighbors, that is, those who have similar values of the attributes should have stronger influence. That's why it is necessary to calculate the distances (or proximity) from the unclassified instance to the *k* nearest instances, in order to take the weights into account. Different formulas can be used for determining the distance, and the general form is Minkowski's formula:

$$Md_{(x,y)} = \left(\sum_{i=1}^{n} | x_i - y_i |^p\right)^{\frac{1}{p}} \quad (1)$$

When $p = 2$ Minkowski distance becomes Euclidean distance:

$$E_{(x,y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

When $p = 1$ Minkowski distance becomes Manhattan ('taxi') distance:

$$d_{(x,y)} = \sum_{i=1}^{n} | x_i - y_i | \tag{3}$$

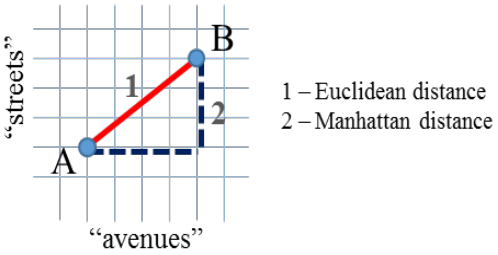

Figure 3. Difference between Euclidean and Manhattan distances

A wide range of different values of numerical attributes can lead to a disproportionate influence of a particular attribute in decision-making. Therefore, the value of the attribute is normalized as follows:

$$x_i^N = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

Often, not all attributes are of equal importance. For this reason, the weights that take these differences into account are used. Regarding the weights, the final distance is calculated according to the following formula:

$$d_i^* = \frac{\sum_{i=1}^{n} \omega_i \cdot d_i}{\sum_{i=1}^{n} \omega_i} \tag{5}$$

d_i – distance of the *i*-th attribute

ω_i – weight (significance) of the *i*-th attribute

After all distances to the *k* nearest neighbors are calculated, it is possible to make the decision on the winning class. The weights of the instances can also be introduced, so that the closer instances have more impact than the more distant ones. The formula used is:

$$N = \sum_{i=1}^{n} \frac{1}{d_i} \tag{6}$$

2.1. MODEL EVALUATION

The evaluation of the model is based on model quality parameters. Most often, the relevant parameters for model assessment are its accuracy, precision, recall and, so called, F-measure of the model. The F-measure is dependent on model's precision and recall as it represents their harmonic mean.
The accuracy of the model (*A*) is calculated as the ratio of the number of accurately classified apartments and the number of incorrectly classified ones:

$$A = \frac{\sum_{i=1}^{n} c_{ii}}{\sum_{i,j=1}^{n} c_{ij}} \tag{7}$$

$C \in \{C_1, .., C_i, .., C_n\}$ – classes

c_{ii} – number of instances that the model correctly classified

$c_{ij}^{\,i \neq j}$ – number of instances for which the model's decision and the actual situation differ

Contingency table is a form used to show the number of instances that are the result of the model's decision. Its elements are shown in *Table 1*.

Table 1: Contingency table

Value	Real	Model
TP_i	(+ C)	(+ C)
FP_i	(- C)	(+ C)
TN_i	(- C)	(- C)
FN_i	(+ C)	(- C)

TP_i – ***TRUE POSITIVES*** – number of instances from class C_i that model classified as C_i

FP_i – ***FALSE POSITIVES*** – number of instances that are not from class C_i but were mistakenly classified as C_i

TN_i – ***TRUE NEGATIVES*** – number of instances that are not from class C_i that model didn't classify as C_i

FN_i – ***FALSE NEGATIVES*** – number of instances from class C_i but were mistakenly not classified as C_i

Using the mentioned values, it is feasible to calculate class C_i precision (π_i), its recall (ρ_i), F-measure for one class (F_i) and model's final weighted F-measure (F):

$$\pi_i = \frac{TP_i}{(TP_i + FP_i)} \;\; ; \;\; \rho_i = \frac{TP_i}{(TP_i + FN_i)} \;\; ; \;\; F_i(\pi_i, \rho_i) = \frac{2 \cdot \pi_i \cdot \rho_i}{\pi_i + \rho_i} \;\; ; \;\; F = \sum_{i=1}^{n} \omega_i \cdot F_i \tag{8}$$

3. EXPERIMENTAL SECTION

The experimental section consists of several segments. The first step are the market analysis and data collection. This segment can be further broken down into finding the data source, deciding on the domain of the problem, choosing the apartments' characteristics that are being considered, collecting data and, finally, forming the dataset suitable for further analysis. An analysis of numerical and categorical attributes was done in the second part. Additionally, adequate statistical parameters were calculated in order to make it easier to see the scope and summary of available data. Third and fourth segments present the results of the classification and regression models, respectfully.

3.1. MARKET ANALYSIS AND DATA COLLECTION

A large number of real estate data can be found publically in advertisements on the Internet. As a data source for this paper, a database of advertised apartments on the website [5] was used, where sellers advertise their real estate. The website contains basic information about the characteristics of apartments that are used as a dataset for defining a model. The data were collected individually, as a stratified sample, by taking one apartment per stratum of the total price (from 10000 to 20000, from 20000 to 30000, ... , from 140000 to 150000 euros).
Dataset comprises of 218 collected apartments. They have next characteristics, with the domain shown in brackets: *Location* (Banovo brdo, Vračar, Sremčica, Ledine, Žarkovo, Vidikovac, Banjica, Braće Jerković, Voždovac, Medaković, Senjak, Dedinje, Cvetkova pijaca, Đeram, Palilula, Dorćol, Stari grad, Bele vode, Zvezdara, Karaburma); *Number of rooms* (studio, 1, 1.5, 2, 2.5, 3, 3+), *Total price (in euros)*, *Area (square meters)*, *Floor* (basement, ground floor, 1, 2, 3, …), *Total floors*, *Top floor* (yes, no, null), *Central heating* (yes, no, null), *Condition (impression based on a photo)* (bad, medium, good, null), *Age* (old, new), *Elevator* (yes, no, not necessary, null), *Clean papers* (yes, no, null), *Balcony* (yes, no), *Basement* (yes, no), *Unit price* (in euros per square meter).
Raw data is usually not convenient for training the model without any prior processing. Hence, some expert knowledge was included. For example, the attribute *Elevator* takes the values "yes" if there is an elevator, "no" if the building does not have an elevator, "null" if the value of this attribute is not known, and the newly created expert value "not necessary" if the apartment is in basement or ground floor. Also, attribute *Condition* was converted to numerical values, so that the available values could be sorted into a sequence and thus be comparable. Accordingly, raw data from the website is mapped into the appropriate set of data that is prepared for defining the model in the next steps.

3.2. STATISTICS OF COLLECTED DATA

The statistics considering attributes of the used dataset are shown in *Table 2* and *Table 3*. The first table contains nominal attributes while the second one depicts numerical attributes.

Table 2: Statistics of collected nominal data

Location	N	%
Banovo brdo	13	6
Vračar	14	6
Sremčica	8	4
Ledine	7	3
Žarkovo	11	5
Vidikovac	9	4
Banjica	11	5
Braće Jerković	12	6
Voždovac	12	6
Medaković	11	5
Senjak	9	4
Dedinje	12	6
Cvetkova pijaca	11	5
Đeram	10	5
Palilula	12	6
Dorćol	12	6
Stari grad	9	4
Bele vode	11	5
Zvezdara	12	6
Karaburma	12	6
∑	218	100

Top floor	N	%
Yes	49	22
No	157	72
Null	12	6
∑	218	100

Central heating	N	%
Yes	96	44
No	110	50
Null	12	6
∑	218	100

Age	N	%
New	46	21
Old	172	79
∑	218	100

N - number of apartments with the corresponding value of the attribute

Clean papers	N	%
yes	162	74
no	25	11
null	31	14
∑	218	100

Balcony	N	%
yes	140	64
no	78	36
∑	218	100

Basement	N	%
yes	66	30
no	152	70
∑	218	100

Elevator	N	%
yes	52	24
no	65	30
not necessary	79	36
null	22	10
∑	218	100

Table 3: Statistics of collected numerical data

Attribute	N	MIN	MAX	MEAN	STDEV	SKEW	KURT	Q1	Q2	Q3	Q3-Q1
Total price [€]	218	14000	195000	77399	38820	0	-1	45588	73250	105000	59413
Area [m^2]	218	14	152	61	28	0.64	0.31	39	59	79	40
Unit price [€/m^2]	218	500	3438	1300	451	1.14	1.83	965	1188	1563	598
Number of rooms	218	0.5	4	2	1.07	0.01	-0.85	1.5	2.5	3	1.5
Floor	218	-1	12	2	3.10	1.41	2.27	0.5	1	3	2.5
Number of floors	218	-1	17	4	3.10	0.92	1.91	2	4	6	4
Condition	218	1	3	1.92	0.76	0.14	-1.27	1	2	2.75	1.75
Percentile	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Total price [€]	14000	29560	39700	50000	61982	73250	85000	97000	118800	135000	195000
Area [m^2]	14	26	36	44	50	59	66	75	83	97	152
Unit price [€/m^2]	500	869	938	1000	1081	1188	1322	1442	1655	1945	3438
Number of rooms	0.5	0.5	1	2	2	2.5	2.5	3	3	4	4
Floor	-1	0.5	0.5	0.5	1	1	2	3	4	5	12
Number of floors	-1	1	2	3	3	4	4	5	6	8	17
Condition	1	1	1	1	2	2	2	2	3	3	3

Particularly important statistical parameters from the table above are the percentile ranks for the attribute *Unit price*. There values were used for defining the boundaries of the future classes within k-Nearest Neighbors classification. As it was decided that there should be no more than six classes, their boundaries were roughly set to the unit prices corresponding to the sextiles. The boundaries were rounded to hundreds of euros. The distribution of unit prices classes can be seen in *Figure 4*.

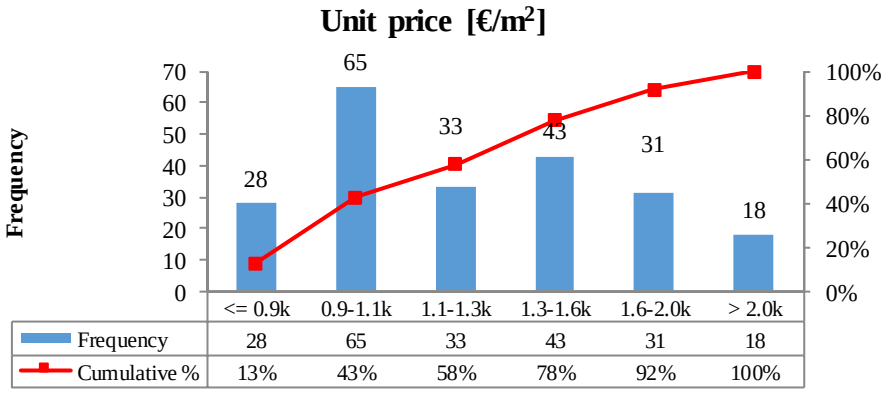

Figure 4. Distribution of unit prices classes

3.3. CLASSIFICATION

The k-Nearest Neighbors classification on the available dataset was performed using the tools from Java-based suite of machine learning software Weka [6]. Firstly, the normalization of all numerical attributes according to formula (4) was performed. The value used for parameter *k* is three. It turned out that it is optimal to use the Manhattan distance shown in formula (3). In addition to this, the proximity i.e. the weights of the instances were taken into account in accordance with formula (6).
Having a relatively small number of apartments dictated that the testing has to be carried out on every instance. The method used for testing was cross-validation and the F-measure is declared as the main quality indicator. The model's quality estimates for KNN classification are shown in *Table 4*.

Table 4: Classification model evaluation

Class	Accuracy	Precision	Recall	F-Measure
<= 0.9k	32.11%	0.190	0.143	0.163
0.9-1.1k		0.354	0.431	0.389
1.1-1.3k		0.286	0.303	0.294
1.3-1.6k		0.270	0.233	0.250
1.6-2.0k		0.290	0.290	0.290
> 2.0k		0.600	0.500	0.545
Weighted average		**0.318**	**0.321**	**0.317**

3.4. REGRESSION

Chapter 3.3 presented a classification model, the purpose of which is to predict the apartments' classes, while the aim of the regression model is to predict the unit price of apartments. The regression model was built using the same parameters and with the same dataset as it was the case with the classification. Estimates of the regression model and unit price prediction for the first five apartments are shown in *Table 5*.

Table 5: Regression model evaluation and the prediction examples

Performance	Value		Apartment	Actual	Predicted	Error
Correlation coefficient	0.5573		1	1042.37	1186.00	143.63
Mean absolute error	289.48		2	1355.93	1411.09	55.16
Root mean square error	376.76		3	1840.00	1423.58	-416.42
Relative absolute error	81.29		4	727.81	1130.97	403.16
Root relative square error	83.31		5	1660.38	1489.69	-170.69

4. DISCUSSION AND CONCLUSIONS

By observing the results of the model evaluation parameters, it can be concluded that the obtained values have low values. The reason for this can be found in many different facts. First of all, the number of downloaded data about the apartments was insufficient. Likewise, it is noticeable that there is a large dispersion of the attribute values for the observed dataset.

Dataset includes some hardly comparable instances. For example, there are old and new apartments, unattractive and luxurious locations, etc. A very important fact to note is that some of the most important characteristics for an apartment are lacking, such as its micro-location, orientation, environment, etc. These attributes are practically impossible to get from the advertisements. Additionally, the shown prices are subjectively determined and are usually ± 20% of the real market values. Also, there is no information as to how much a seller is hurrying, or whether the given price is perhaps a liquidation value, when the seller is forced to sell the apartment in a short period of time.

The classification model gives better results when the number of classes is reduced, but in that case, the question of the usability of the model arises. The k-Nearest Neighbors method requires strict boundaries between classes, which does not correspond to the nature of this problem. A possible way of overcoming this deficiency would be the fuzzification of the selected classes. As for the regression model, on this set it returnes unit prices with a large deviation, i.e. it does not provide the expected performance for the above reasons.

In order to improve the model, it would be necessary to develop a system for collecting a large amount of accurate, precise and up-to-date data that would allow one such model to be applied in practice. By collecting data on very similar apartments, the model would be reduced to the automation of a widely-used comparative method for assessing the market value of the real estate. This would be the proposed direction for further research.

When defining a model, it is very important to limit it to the appropriate domain of the problem. The number of different attributes and potential attribute values should be adjusted to the available data. Non-informative attributes should be removed which would lead to reducing noise as much as possible. If the model included those very informative but hardly accessible features, there would be a significant increase in its capacity and this would favor the justification of applying this approach.

Acknowledgements

The authors would like to thank the Ministry of Education, Science and Technological Development of the Republic of Serbia for the financial support through the projects III 47014 and TR 36020.

REFERENCES

[1] Rulebook on National Standards, Code of Ethics, and Rules of Professional Conduct for Licensed Valuers, *RS Official Gazette*, № 70/2017

[2] Han, J., Kamber, M., Pei, J.: *Data Mining Concepts and techniques (3rd ed.)*, Elsevier, 2011.

[3] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 1996., vol. 17, № 3, p.p. 37-54.
[4] Sumathi, S., Sivanandam, S. N.: *Introduction to Data Mining and its applications*, Springer, 2006.
[5] https://imovina.net/, download on 4th February 2019
[6] https://www.cs.waikato.ac.nz/ml/weka/, download on 12th February 2019

PROCENA TRŽIŠNE VREDNOSTI STANA METODOM k-NAJBLIŽIH SUSEDA

Rezime: *Tržišnu vrednost stanova, kao što sama reč govori, definišu prodavci i kupci kroz ponudu i tražnju, koje u osnovi i čine samo tržište. Sagledavanje velikog broja faktora uticaja na cenu nepokretnosti nije nimalo lak posao. Formiranje cena zavisi kako od karakteristika stana, tako i od sistema vrednosti kupaca. „Zašto bih ja za istu ili sličnu stvar platio veći iznos nego što je neko drugi platio u neposrednoj prošlosti" jeste osnovno pitanje koje racionalan kupac postavlja sebi. Ova činjenica dovodi do zaključka da je potrebno znati karakteristike i cene nepokretnosti koje su oglašene ili prometovane u neposrednoj prošlosti u bliskom okruženju. Komparativni način razmišljanja kupca je osnovni uslov i princip za definisanje jednog ovakvog modela. Ovo je neophodan, ali ne i dovoljan uslov. Modeli bazirani na algoritmima mašinskog učenja, kao što je i k-najbližih suseda, podrazumevaju poznavanje nešto veće količine kvalitetnih podataka, kako bi doneti zaključci bili pouzdani, tačni i precizni.*

Ključne reči: *procena tržišne vrednosti, tržište stanova, zaključivanje iz podataka, metoda k-najbližih suseda*