THE USE OF REGRESSION TREES TO THE ANALYSIS OF REAL ESTATE MARKET OF HOUSING [1]

Elżbieta Jasińska, **Edward Preweda**
AGH University of Science and Technology, **Poland**

ABSTRACT

Housing property can be described using a number of traits, some of which turn out to be difficult to express in numerical form. You can also use parameters which are not subject to explicit quantification. Belong to it selected location or membership of an exclusive building project, or opposed to a specific buyer reluctance of the complex. The attributes relevant to buyers may be disregarded just because of problems with their recognition in shaping market prices. A multitude of real estate tenancy causes difficulties with their comparison. If you make this scaling problem adopted features such analysis may not be sufficient. An attempt to include as many information may be restricted by the form of analysis.

The undoubted advantage of the classification tree is the ability to delineate the characteristics of a quantitative (continuous or discrete) on a par with those of a qualitative nature (from the nominal scale after an ordinal). They allow an assessment of the impact of both the characteristics of qualitative, as well as a quantitative variable quantitative, without having to specify an arbitrary numeric values for the variables on the nature of the quality. This option allows you to include attributes such as subdivision, street name.

Among the many methods of data analysis, it is worth noting the use of regression task for classification trees, so that the designation of property prices was possible based on the allocation to one of the nodes in the schema that you created earlier. The proposed method allows for the examination of the impact on the predictors of the dependent variable, as well as on the distribution of the existing set of homogeneous groups in terms of price.

Keywords: attributes of real estate, real estate analysis, regression trees

INTRODUCTION

Information on the preferences of the buyers and the determination of market trends is especially valuable to participants in the real estate market. Many investors and developers are interested in studies and reports prepared by specialized consulting company, or works with the valuers. Banks also expect confirmation of real estate prices, before granting a mortgage. Statistical analysis of the market lets clearly and concisely characterized in it trends and phenomena, so that experts can draw

[1] This work is financed from funds for science realized at AGH University of Science and Technology, allocated for the year 2013

conclusions and make substantive comparisons and generalizations designed to detect regularity. It is worth to consider whether all the parameters are taken into account and as far as within the eg. the selected area or street, price differences will be resulted from the attributes other than the position or location. A way of describing each attribute category, beginning their counts, and assigned them to numeric values should arise directly from the nature of the market. However, in both these issues it is hard to look for the single position of the whole environment [3]. Therefore, the introduction of classification trees would at least for skipping to assign numeric values and work on the same names. The new proposal extends the existing research through concurrent consideration of the qualitative and quantitative characteristics (without having to assign numeric values). Such a study was conducted by the authors using regression trees, for example, the C&RT model (Classification and Regression Tree).
It also uses the C&RT tree at work [2] concerning the characteristics of impact studies on the importance of given property. In this monograph , it was considered that the most important impact on the development of price movements, it has mainly over the city in which the property is located, regardless of other characteristics. Use this algorithm for classification tasks is common, while the regression problem is no longer as common. Also literature concerning regression approach is poorer [1].

RESEARCH METHODOLOGY

The purpose of the research was to assess the applicability of regression trees for determining property prices, while utilizing the characteristics of qualitative and quantitative in nature. Particularly important was the designation of the unit price volatility based on the specific road, or at the very least.

C&RT TREE MODEL

The behavior of the algorithm C&RT (Classification and correct Regression Tree) is a binary recursive distribution group objects, each test is based on one variable [1]. If it is possible to select such divisions, so that the variance in the end nodes is zero, it says that the pollution of the node is the minimum and maximum uniformity. In practice, however, it is very difficult. The vector data clearly "descends" the path from the root to a leaf.
The C&RT algorithm aim is to divide the data in a way which allow for the separation of high and low values of the dependent variable, which means that in a properly built model one dedicated node values are higher, and the second lower than the values in the node. Finding a compromise between the size of the tree, and as little growth prediction error is possible thanks to the principle of one standard deviation (1 SE rule). The authors of the C&RT algorithm suggest as right-sized tree to choose which test costs do not exceed the cross-cross test minimum plus the value of the standard deviation of these costs[1].
On the basis of the creating process of a model, it is possible to systematize these predictors of which have substantial predictive power relative to the dependent variable. The validity of predictors is the inverse of the sum of the cost of resubstitution (replace) on all nodes in the tree created for a given variable. It is expressed on a scale of 0-1 (scaled so that its maximum is 1), which can serve as analogy to the correlation coefficients, however, ranked the validity of it cannot be determined whether the feature affects positively or negatively on the value of the variable. It is possible to find that predictor, which occurred not as a criterion for the final distribution of the tree, even

though it has a high place in the ranking. It is possible because such an attribute during most divisions, was second to the possibility of reducing the variance in knots-kin. Despite that, ultimately, was not placed on the graph, its predictive capability is greater than this attribute, who "took advantage of the predictive power" in the first divisions of the tree and then was not relevant for the further segmentation.

RESEARCH MATERIAL

Information about more than 1,100 real estate sold in the past 3 years, then their characteristics updated on the date August 1, 2012. Database of real estate premises intended for housing is divided into 14 groups, according to districts. Two districts were omitted due to a small number of transactions. For each of the transactions was collected the information presented in table 1. The first step was to create a tree model for C&RT the entire area regression chases Krakow

Table.1. Description of a data set from the market dwellings, based on [4]

Feature	Category	Description	Comments
District	The name of the district	Territorial affiliation to one of the mentioned districts of the city of Krakow	Feature expressed in nominal scale. Impossible to ranking prior to analysis. Due to the its nature.
Street	The name of the street	Membership features by property address	Feature expressed in nominal scale. Impossible to rank before performing any tests.
The Surface	0 ÷ 196	The surface of the place expressed in [m2]	Feature expressed in quantitative scale continuous
The Surroundings	Adverse Benefit Very Beneficial	Neighborhood, its location relative to commercial venues, sports and green areas	Feature in the ordinal scale, it is possible to assign numeric values
Standard	Low Medium High	Type of fenestration, floors	Feature in the ordinal scale, it is possible to assign numeric values
Transport access	Negative Average Positive Very positive	The number of available communication lines and the distance from the bus stop	Feature in the ordinal scale, it is possible to assign numeric values
Technical usage	0 ÷ 100	The consumption of functional premises expressed as a percentage	Feature expressed by numerical values
Arrangement of the premises	Negative Average Positive Very positive	Arrangement of rooms in the premises, non-transitive, narrow passages,	Feature in the ordinal scale, it is possible to assign numeric values
Floor	0 ÷ 10	The location of the premises in the building (ground floor adopted as "0")	Quantitative trait
Year of construction	1910 ÷ 2008		Quantitative trait
Unit price	0 ÷ 11 000	Price per 1 [m2] place sold	Feature of a continuous quantity

DISCUSSION OF THE RESULTS

On the basis of such data, regression tree was prepared. Calculations was done by using Statistica, therefore it was not possible to take into account the characteristics of the street, which contains more than 200 values (the program could not process the number of variables). The schema created the tree shown in Figure 1, this schema meets the criterion of a single error. The database can be efficiently divided into 12 subsets of, there are nodes-list numbers: 4, 5, 12, 13, 16, 18, 19, 26, 27, 28, 29. In the first place were the most prestigious around, including the most expensive, the Old Town and Zwierzyniec (nodes 5 and 4), it should be note that only node nr 4 is homogeneous in terms of belonging to the district, node 5 have mixed properties.

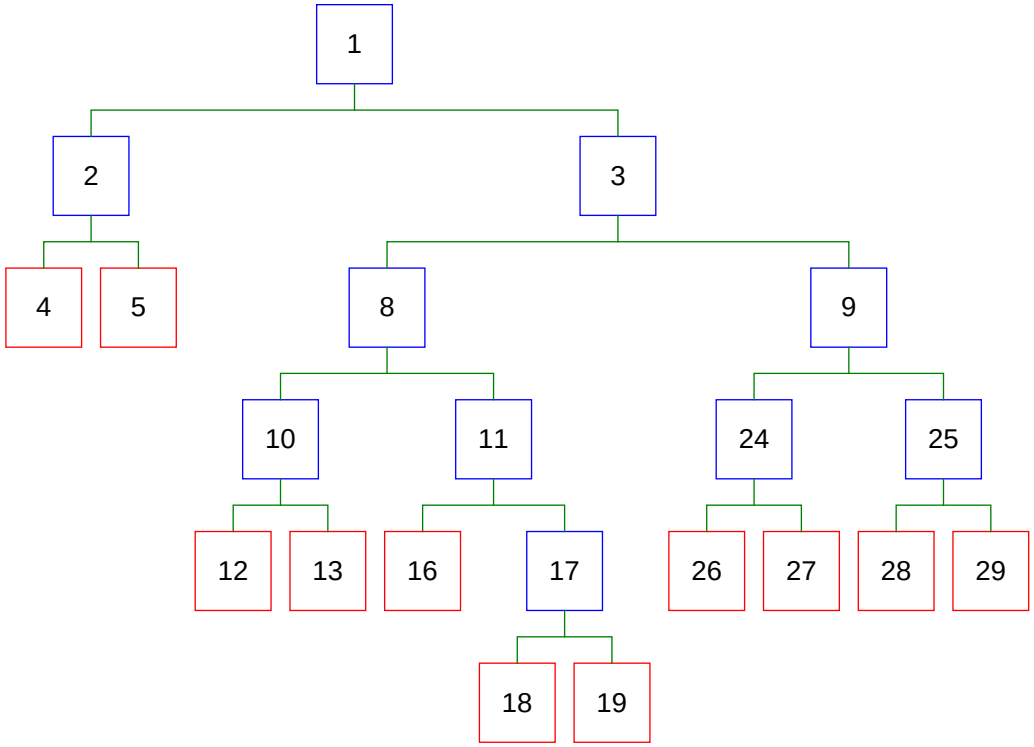


Figure 1. The schema tree for the whole city

In the process of analysis of the validity of the chart was created all attributes, which shows that this subdivision is the most important factor affecting the price, so if they accumulate in themselves generally defining features of the property. Since it is the main criterion, further analysis should be carried out inside the neighborhoods.
To this aim, the Old Town district was chosen as the most representative and reiterated the create schema, by adopting the following parameters for the tree C&RT:

- Shared collection of counts 167 real estate,
- Predictor: unit price,
- The quality characteristics: Street, Neighborhood, Standard, Transport Access, Location of the premises,

- Crop according to variance-if either of the nodes of the descendants of the variance is not decreasing, it's such a partition is deleted in the process of trimming the trees,
- The maximum number of nodes: 1000,
- The minimum number in a node: 16.

Right-sized tree has been selected on the basis of the cost of 10-cross test. Best predictive capacity tree is considered the basis for its price-setting attributes and the relationships between them, the appropriate tree shown in Figure 2.

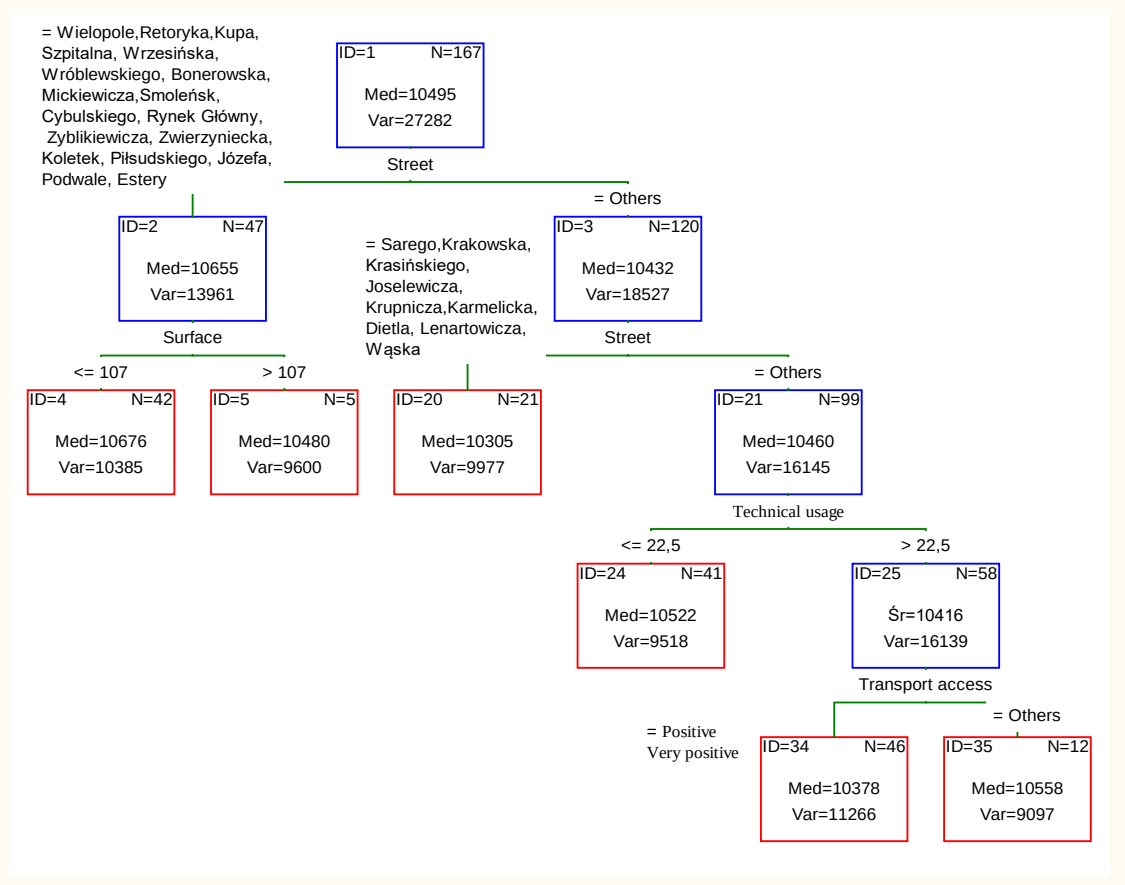


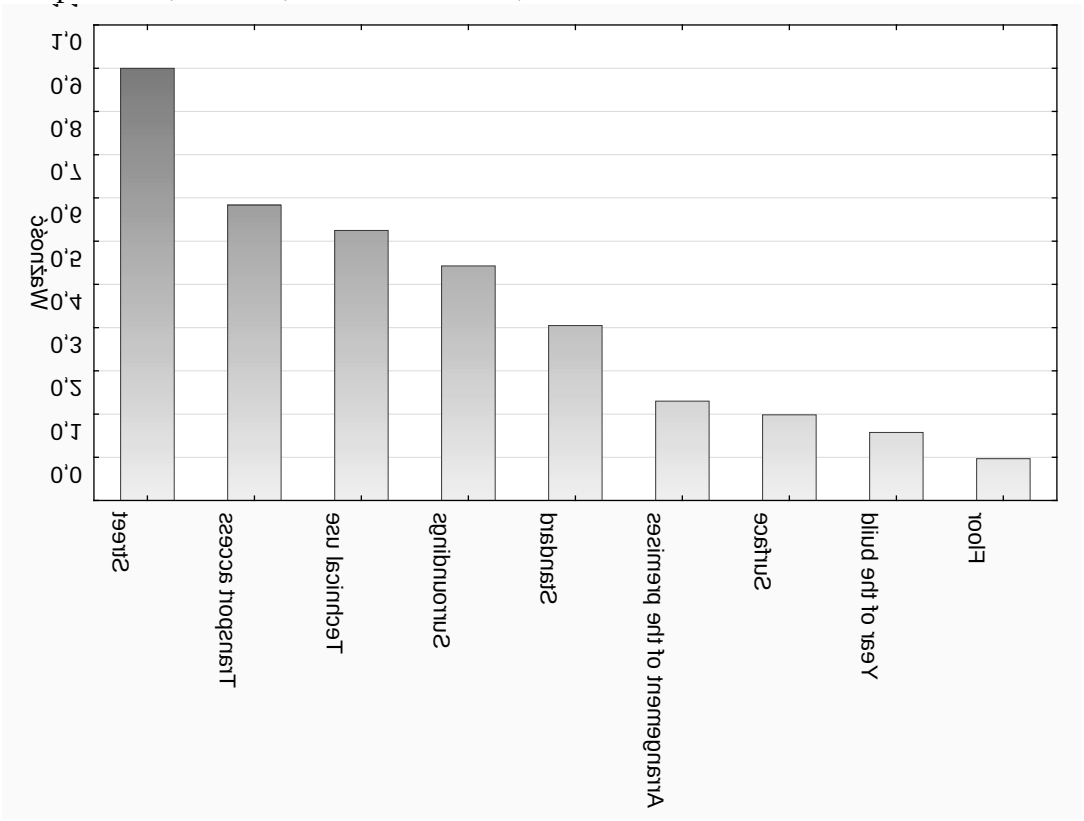
Figure 2. C&RT for Old Town district

On the basis of the schema shown in Figure 2, expert can estimate the average unit price obtained on the basis of divisions, in addition to this price will be quoted with estimated error of its designation, which may be presented in the form of conditional sentences:

- If the property is situated at:Wielopole, Retoryka, Kupa, Szpitalna, Wrzesińska, Wróblewskiego, Bonerowska, Mickiewicza, Smoleńsk, Cybulskiego, Rynek Głowny, Zyblikiewicza, and its surface does not exceed 107 square meters, its average unit prices shall be 10 676 PLN ± 102 PLN (node id=4)
- Conversely, if the midwife is on the streets: Sarego, Krakowska, Krasińskiego, Joselewicza, Krupnicza, Karmelicka, Dietla, Lenartowicza, Waska regardless of the other features, you can determine the mean price on 10305 PLN ± 100PLN (node id=20)

For the remaining nodes relevant information was the availability and technical use of transport. In parallel with the creation of tree and subsequent bands can carry out an assessment of the impact of the evolution of attributes average unit price, ranking the

characteristics shown in the figure 3, according to which the most important feature is the street, later: Transport access, Technical use, Surroundings, Standard, Arrangement of the premises, Surface, Year of the build, Floor.



CONCLUSIONS

Classification and regression trees can be used as a tool to check or estimate the price of the real estate. Every district can have different division rules, but for statistical experts it is not so difficult to create C&RT thee for each one. Of course it can be create as a big one for the whole city, however it is not possible to include every street in it. That's why it's much comfortably to make such divisions based on the attributes characteristic for every enclave. This paper present that offered model can help to estimate unit price with standard deviation on the level of view percent. In parallel with the tree creation the ranking of predictor is preparing, so it can be rated which of attributes are the more or less important in price changings.

REFERENCES

[1] Hand D., Mannila H., Smyth P.: Principles of Data Mining, MIT Press, 2001
[2] Jasińska E.: Wybrane metody statystyczne w analizie rynku nieruchomości (Chosen Statistical Method In Real Estate Market Analysis), UWND, AGH, 2012
[3] Jasińska E., Preweda E.: Methods Of Selecting Factors In The Analysis Of The Real Estates Market. Geodezja, T.12, z. 1, AGH, Kraków, 2006
[4] Czaja J., Parzych P.: The estimation of real estates' market value in the aspect of International Valuation Standards (Szacowanie rynkowej wartości nieruchomości w aspekcie Międzynarodowych Standardów Wyceny), Krakow, 2007