ADVANCED COMPUTING SESSION

AN ALGORITHM FOR THE MULTIDIMENSIONAL ANALYSIS OF THE OVERESTIMATE AND UNDERESTIMATE OF PROPERTY RENTAL VALUE

Mladen Mijatović

Singidunum University,
Belgrade, Serbia

Abstract:
Multidimensional attributes related to a real estate property interfere with the cognitive ability of potential real estate property user to process all available information in order to determine the real rental market rate and make a fully informed decision. Although there are various web platforms for property leasing, it is difficult for a potential user to find two equivalent real estate properties and thus make relevant comparisons. This paper proposes a potential solution for the rental price estimate of a property to be leased, as well as scaling and comparison of its overestimated or underestimated value to all other available properties. This example is developed on the basis of real estate market in Sarajevo. Random Forest algorithm was used for the development of regression model. The principle aim of this paper is to offer an applicative module, based on Data Science methodology, for automatic scaling and the comparison of property overestimated or underestimated rental rates to other residential buildings and taking in consideration multidimensionality of their physical and geolocation attributes.

Keywords:
Real Estate, Price Estimation, Random Forest Regression.

INTRODUCTION

Information as a resource has become an important commodity in every business and its importance is reflected in aiding, that is enabling quality decision making for a decision maker. Unequal availability of information to service providers and service users, also known as information asymmetry, is also found in a real estate market. Information asymmetry can be seen in different economic transactions mostly in situations when one party on the market possesses greater material knowledge than the other and where the better-informed party is a service provider [1]. The importance of information in a decision-making process increases with the value of goods / services and this is particularly pronounced in a real estate market. Also, the decision-making process becomes more

Correspondence:
Mladen Mijatović

e-mail:
mladen.mijatovic.19@singimail.rs

DOI: 10.15308/Sinteza-2020-268-274

complexed with the increased attributes dimensionality of products / services and such decisions are made more slowly by the users compared to decisions made when purchasing low-value products/services. Multidimensional attributes related to a real estate property interfere with the cognitive ability of potential estate property user to process all available information in order to determine the real market rental rate and make a fully informed decision. When researching properties, users usually make their estimates based on personal experience, interest and sometimes even include emotional aspect. Although there are various web platforms for property leasing, it is difficult for a potential user to find two equivalent real estate properties and thus make relevant comparisons. In regard to the described problem, the subject of this paper is an analysis and modeling of rental value taking in consideration geolocation and physical attributes of the property. This paper proposes a potential solution for the rental price estimate of a property to be leased, as well as scaling and comparison of its overestimated or underestimated value to all other available properties. This example is developed on the basis of real estate market in Sarajevo. Random forest algorithm was used for the development of regression model. The principle aim of this paper is to offer an applicative module, based on Data Science methodology, for automatic scaling and the comparison of rental properties and their overestimated or underestimated value rates to other residential buildings and taking in consideration multidimensionality of their physical and geolocation attributes. Model development consists of two steps. The first step is to predict a real estate property value using random forest algorithm, while the second step is to conduct the scaling of the overestimate or underestimate of the property value based on the rating of the obtained residual values in regression and their discretization.

1. PREVIOUS RESEARCH

Recently a lot of research has been conducted on housing price forecasts either for sale or leasing (houses and apartments). Models for property price estimate were developed through different regression algorithms, using either traditional statistical algorithms, such as multiple linear regression or neural networks and deep learning algorithms. So, the group of authors [2] has done a research on the predictive performance of the random forest algorithm in comparison to commonly used hedonic models based on multiple regression for the prediction of real estate property prices in the city of Ljubljana, the capital of Slovenia. All performance measures (R2 values, sales ratios, mean average percentage error (MAPE), coefficient of dispersion (COD)) revealed significantly better results for predictions obtained by the random forest method. In a paper on a similar topic [3] it was proven that compared to linear regression model, random forests model can better capture hidden nonlinear relations between the price and features of a real estate property and in overall give a better estimate. The absence of multiple linear regression analysis in property price estimation is confirmed in paper [4], where authors suggest using hierarchical linear model (HLM) instead of standard linear model. Joshua Gallin [5] used standard error-correction models and long-horizon regression models to examine how well the rent–price ratio predicts future changes in real rents and prices and found evidence that the rent–price ratio helps predict changes in real prices over 4-year periods, but that the rent–price ratio has little predictive power. In one of the papers [6] author examined the determinants that significantly influence apartment prices that are located within housing estates of Nairobi metropolitan area. Author used multiple regression analysis and the findings indicated that land value and size of the apartments had a significant influence on apartment pricing. A machine learning algorithms trained on big data have a great potential for prediction of property price estimate compared to small datasets [7]. Also, artificial neural network (ANN) has better predictive accuracy in property valuation in comparison to estimations made using traditional hedonic pricing model [8].

2. RANDOM FOREST REGRESSION

Random forest algorithm is becoming increasingly popular in many scientific fields because it can cope with so called "small n large p" problems, complex interactions as well as highly correlated predictor variables [9]. Due to its performances, random forest is more robust compared to multiple linear regression which complies with statistical assumptions such as homogeneity of variance, lack of multicollinearity, linearity of variables, etc. Random forest algorithm belongs to the group of algorithms for supervised machine learning.

Random forest algorithm can be used either for categoric response variable, when algorithm solves “classification” problem or continuous response variable, when algorithm deals with “regression” problem. Similar to that, predictor variable can be either continuous or categorical.

From an application standpoint, random forest algorithm is appealing because it [10]:

- naturally handle both regression and classification problems;
- are relatively fast to train and to predict;
- depend only on one or two tuning parameters;
- have a built-in estimate of generalization error;
- can be used directly for high-dimensional problems;
- can easily be implemented in parallel.

Statistically, random forest algorithm is appealing because of the additional features it provide, such as [10]:

- measures of variable importance;
- differential class weighting;
- missing value imputation;
- visualization;
- outlier detection;
- unsupervised learning.

Random forest algorithm is a type of additive model that makes value predictions by combining decisions from a sequence of base models.

Formally we can write this type of models as [11]:

$$g(x) = f0(x) + f1(x) + f2(x) + \ldots \quad (1)$$

where the final model g is the sum of simple base models fi. Here, each base model is a simple decision tree algorithm. This technique of using multiple regression or classification models to obtain better predictive performance is called model ensembling. In random forest algorithm all the base models are constructed independently using a different subsample of the data.

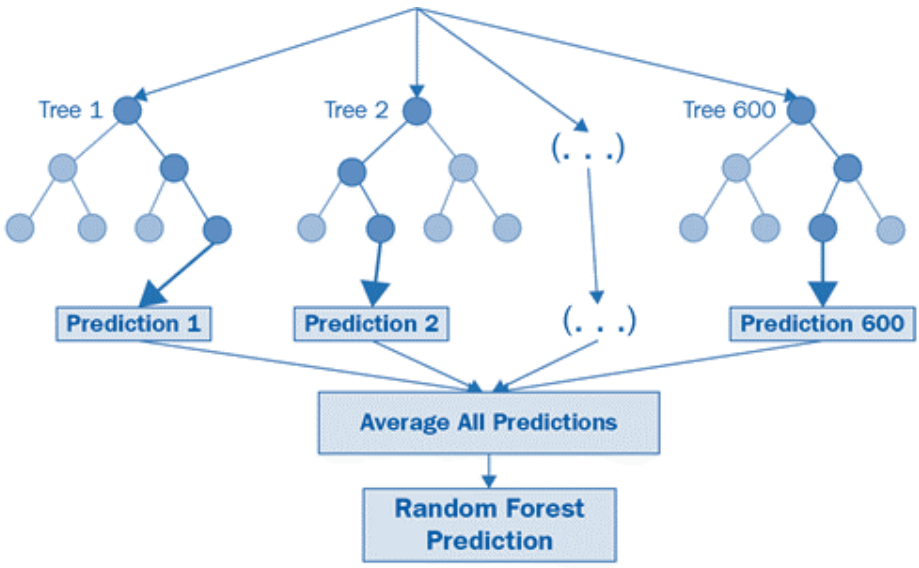

Fig 1. Random Forest Structure

Random Forests is a multipurpose tool, applicable to both regression and classification problems, including multiclass classification. It give an internal estimate of generalization error so cross-validation is unnecessary. It can be tuned, but often work quite well with default tuning parameters. [10]. Random forest method is an effective tool in prediction. Because of the *Law of Large Numbers*, it do not overfit. Injecting the right kind of randomness makes it accurate classifiers and regressors [12].

3. MODEL DEVELOPMENT

Data Collection

For the purpose of developing multidimensional analysis model to determine overestimated and underestimated rental price, data were collected from two sources. The first dataset was collected by web scraping of a website that advertises real estate properties selling and renting in larger cities in Bosnia and Herzegovina [13], and the data extracted refer to available real estate properties and their features such as:

- Type of a real estate property (apartment, house, office space)
- Address
- Price
- Area
- Number of rooms
- Number of bathrooms
- Year of construction
- Adaptations
- Floor
- Amenities (storage room, parking space, garage, AC, Internet access, elevator, etc.)

Data on the website were scraped in the R programming language with "rvest 0.3.4" library [14]. In total 629 rental properties were collected that were located in the city of Sarajevo.

The second dataset was obtained by calling Google Maps API [15] and gathering different geolocation data for every real estate property, as well as the distance of each real estate property from the set reference points:

- The geographic coordinates of a real estate property
- The altitude of a real estate property
- The number of universities in a 2,000-meter radius
- The number of primary and secondary schools and other educational centers (other than universities) in a 700-meter radius
- The number of shopping malls in a 500-meter radius
- The distance of a real estate property from reference points (from Bascarsija, from Vjecna Vatra Monument and from University Campus in Sarajevo).

Places API, Geocoding API, Elevation API and Distance Matrix API were used.

Data Preprocessing

There were some gaps in the data collected that may influence the quality of the final model. For example, from the total of 629 real estate properties, 166 were business offices that were not in the scope of modeling and were removed from the dataset. A smaller number of real estate properties that were located 20 km outside of Sarajevo were also removed. Also, there were 122 real estate properties that did not have quoted price (price on application) and as such could not be used for model review and were removed from the dataset. The final dataset consisted of 343 real estate properties (houses = 61, apartments = 282) located in Sarajevo.

Text data type (string) was converted by parsing and modified into usable modeling attributes. Data obtained with Google Maps API were checked and if necessary, cleared from illogical objects, that were aggregated on each real estate property and using datasets connected to properties.

After we connected datasets, we had 71 attributes that were related to different physical features of real estate properties and their geolocation, as well as specific characteristics of their surroundings.

Feature Engineering

A one hot encoding procedure was conducted for all categorical variables to obtain binary variables usable for random forest algorithm. The date attribute *year of construction* was set by calculating the age of the building as the difference between the current year and listed year of construction (age of the building = 2020 – year of construction).

Outlier Detection

Presence of extreme values in the data was analyzed using Cook's distance method. Cook's Distance (D_i) is an influence measure based on the difference between the regression parameter estimates and what they become if the i-th data point is deleted. There are numerical rules for assessing Cook's distance but the rules tend to be rough guidelines, and textbook authors differ in their advice [16]. Real estate properties that had Cook's distance 4 times bigger than arithmetic average of the distance were considered extreme. Fig 2 represents cases that had extreme values of certain variables (above the red line). In such cases, detected extreme values of some attributes are replaced with missing values.

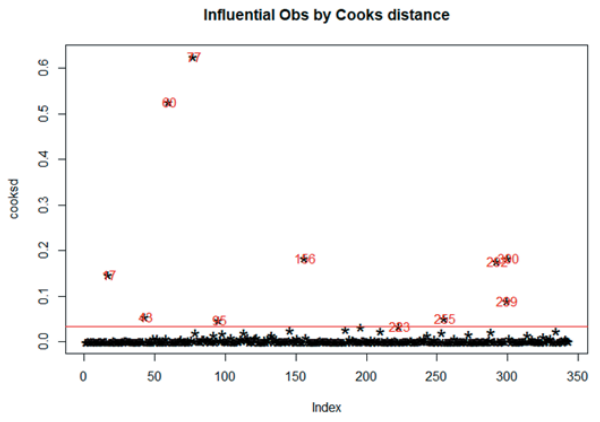

Fig 2. Influence measure with Cook's Distance

Transformation Data

In order to normalize values, all numerical attributes were transformed before the modeling. As some attributes had negative values, we decided to use Yeo-Johnson transformation. If values are strictly positive, then the Yeo-Johnson transformation is the same as the Power Cox transformation [17]. An example of transformation of one attribute is shown in Figures 3 and 4.

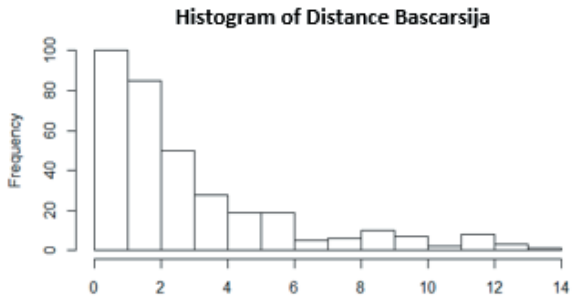

Fig 3. The distribution of the distance values of a real estate property from Bascarsija before Yeo-Johnson transformation

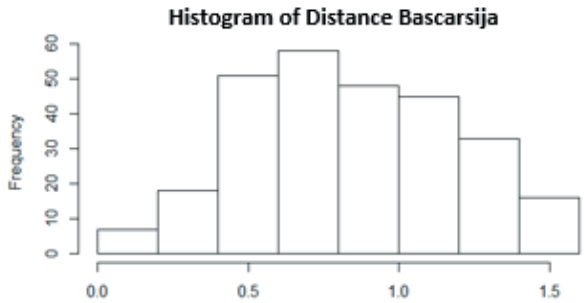

Fig 4. The distribution of distance values of real estate property from Bascarsija afterYeo-Johnson transformation

Model Training

The "caret 6.0-85" R library was used to run the Random forest algorithm [18]. The process of training a model for the rental price estimate, on the basis of random forest algorithm, was conducted on 80% (n = 276) random observations. Random Forest is initiated with n=130 of trees to grow. The 10-fold cross-validation procedure is repeated 3 times. Root Mean Square Error (RMSE) parameter was used for the selection of optimal model. Table 1 shows resampling results across random forest parameters with cross-validation procedure. The final number of variables for splitting a node, and that were used for the model, was mtry = 26 and that number represents the final optimal value with the smallest Root mean squared error (RMSE=1.042) and parameter R-squared of 0.69542.

Figure 6 shows top 12 variables that are the most important in determining variables in rental price estimate in the model. Property area has by far the biggest influence on property value. The number of rooms in the property is the second important variables for rental price estimate. Distance of property from Bascarsija is also important variable for determining the rental price, but also the distance from Vjecna Vatra Monument as the second reference point. The number of universities in 2,000-meter radius from the property is also one of top 12 variables that determine property value. Important variables are number of bedrooms, altitude as well as the floor of the property.

mtry	RMSE	Rsquared	MAE
2	1.198122	0.6673015	0.9653732
5	1.089086	0.6896604	0.8731576
8	1.067655	0.6929206	0.8517182
12	1.059462	0.6916996	0.8415689
15	1.058859	0.6887428	0.8430591
19	1.048629	0.6942377	0.8321179
22	1.048378	0.6924978	0.8290187
26	1.042050	0.6954187	0.8279170
29	1.055678	0.6846355	0.8314533
32	1.051792	0.6870910	0.8287001
36	1.042958	0.6924899	0.8196449
39	1.053497	0.6845814	0.8260079
43	1.058534	0.6806108	0.8248516
46	1.046557	0.6879252	0.8196265
50	1.064484	0.6769833	0.8322964

Fig 5. Resampling results across tuning parameters

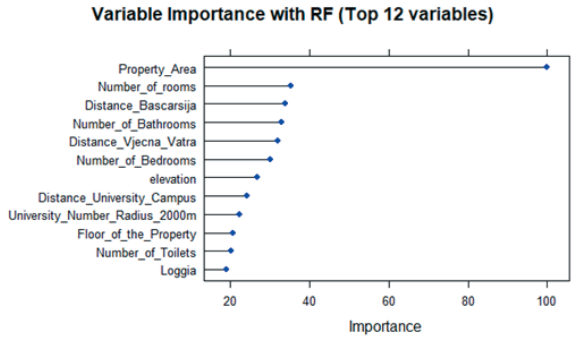

Fig 6. Variable Importance with Random Forest (Model Training)

Model Evaluation

Model evaluation was conducted on the other 20% randomly selected dataset observations. A prediction of rental price was made on a test dataset during random forest model training. A value of Root Mean Square Error obtained was 1.0566 between actual prices and predicted prices with model with training dataset. Scatter plot on Figure 7 shows the relationship of actual prices and predicted prices with the model.

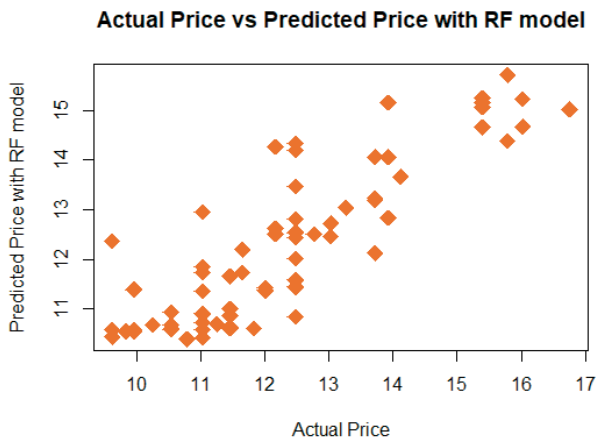

Fig 7. Relationship of actual prices and predicted prices with Random Forest model

4. THE APPLICATION OF MODEL FOR THE SCALING OF OVERESTIMATED AND UNDERESTIMATED RENTAL PROPERTIES

After the prediction of rental price estimate, the next step is normalization and discretization of residuals. In the context of this research, residuals represent the differences between rental actual prices and predicted prices with the model, that is:

$$e = y - \hat{y} \tag{2}$$

where:

e – residual

y - actual price

ŷ – predicted price with the model

Therefore, predicted price with the model $\hat{y}$ represents the value determined by multidimensional influence of variables in the model which are responsible for determining the rental price. Positive residual value of the property shows us the increase of actual price (overestimated) in comparison to expected price determined by the model, while negative residual value shows the decrease of actual price (underestimated) in comparison to expected price determined by the model. Realized residual values are further normalized by Z-score in order to compare certain properties with others in the data base. Finally, in order to determine whether the rental price of property is overestimated or underestimated, depending on its characteristics and compared to other properties and their characteristics, Z-score is discretized in 5 groups ranging from "Very good price" to "High price". Figure 8 shows an example of the scaling of overestimated and underestimated rental value.

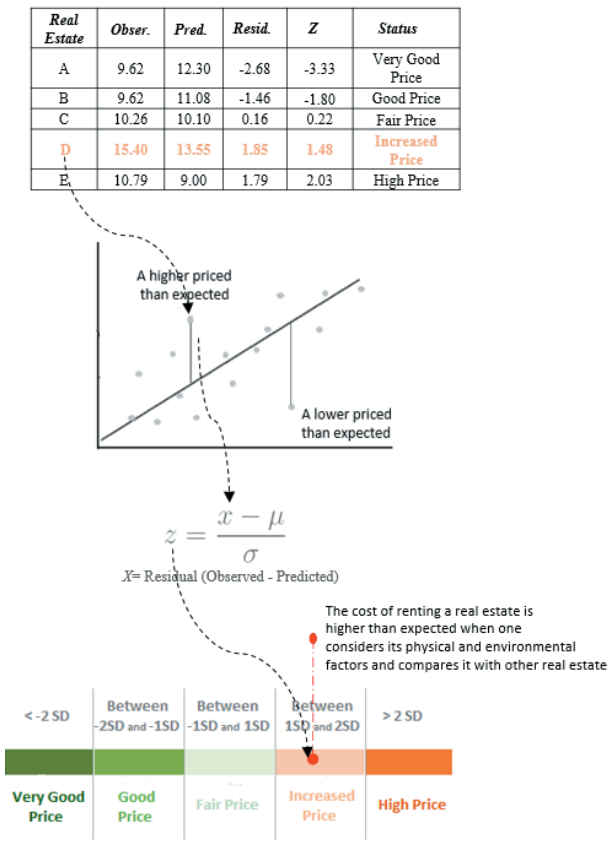

Fig 8. A scaling example of the over- and underestimated rental value

5. CONTRIBUTION OF RESEARCH, LIMITATIONS AND FUTURE RESEARCH

Implementation of presented algorithm on web platforms, which act as intermediaries between real estate service providers and users, will help users (property seekers) with cognitive reduction of the amount of in-

formation as an alternative to filtering of different attributes (location, number of rooms, floor, etc.) in order to make comparison and determine property value. One of the limitations is the development of the proposed solution on a relatively small dataset, which influenced model performances and gave a slightly larger estimation error (RMSE) in the rental value of property. Different researches have shown that the model performances can significantly increase. In future research we should have at our disposal large dataset that would enable us to apply deep learning algorithm. In order to improve model performance in estimation of property value, we should include additional variables that would further provide an explanation to the variability of property prices.

REFERENCES

[1] M. Parkin, Economics (11th Edition), University of Western Ontario, 2014.

[2] M. Čeh, M. Kilibarda, A.Lisec, B.Bajat, Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments, International Journal of Geo-Information, 2018.

[3] C. Wang, H. Wu, A new machine learning approach to house price estimation, BISKA- International Open Access Journals, 2018.

[4] I. Arribas, F. Garcia, F. Guijarro, J. Oliver, R. Tamošiūnienė, Mass appraisal of residential real estate using multilevel modelling, International Journal of Strategic Property Management, Volume 20, p. 77-87, 2016.

[5] J. Gallin, The Long-Run Relationship Between House Prices and Rents, Real Estate Economics, 2008.

[6] J.N. Ndegwa, Determinants of Apartment Prices within Housing Estates of Nairobi Metropolitan Area, International Journal of Economics and Finance; Vol. 10, No. 6, 2018.

[7] J. I. Pérez-Rave, J. C. Correa-Morales, F. González-Echavarría, A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes, Journal of Property Research, p. 59-96, 2019.

[8] R. B. Abidoye, A.P.C. Chan, Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network, Pacific Rim Property Research Journal, Volume 24, p. 71-83, 2018.

[9] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis Conditional variable importance for random forests, BMC Bioinformatics, 2008.

[10] A. Cutler, D. R. Cutler, J. R. Stevens, Random Forest, C. Zhang, Y. Ma (Editors), Ensemble Machine Learning (Methods and Applications), p. 157-175, Springe, 2012

[11] Turi Machine Learning Platform User Guide. [Online]. Available: https://turi.com/learn/userguide/index.html

[12] L. Breiman, Random Forests, Kluwer Academic Publishers. Manufactured in The Netherlands ,2001.

[13] Company "Prostor" Ltd. Sarajevo, [Online]. Available: www.prostor.ba

[14] H. Wickham, rvest library for scrape information from web pages, [Online]. Available: http://rvest.tidyverse.org/

[15] Google Cloud, Google Maps Platform, [Online]. Available: https://cloud.google.com/maps-platform

[16] B. McDonald, A Teaching Note on Cook's Distance – A Guideline, Institute of Information and Mathematical Science, Massey University at Albany, Auckland, N.Z., 2002.

[17] S. Weisberg, Yeo-Johnson Power Transformations, Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108-6042, 2001.

[18] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z Mayer, B. Kenkel, R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt, Classification and Regression Training, [Online]. Available: https://cran.r-project.org/web/packages/caret/caret.pdf