CONSTRUCTION OF TRAINING DATA FOR PRICE PREDICTION OF A REAL ESTATE FROM INTERNET ADS

Mladen Vidović*,
Ivan Radosavljević,
Aleksandra Mitrović,
Zora Konjović

Singidunum University,
Belgrade, Serbia

Abstract:

The paper presents a model for constructing a data set aimed at predicting a price of a real estate (houses and flats) from the standard Internet ads. The model for predicting a real estate price includes, in addition to standard real estate's features (area, number of bedrooms, etc.) appearing in ad, attractiveness of a real estate location as well as information on some additional interior facilities (e.g., refrigerator, dish-washing machine, stove, etc.). The proposed training set construction model uses OpenStreetMap's Overpass API for determining attractiveness of a real estate's location, and a convolution neural network for detecting interior facilities from real estate photos.

Keywords:

real estate, advertising, price, prediction.

1. INTRODUCTION

Purchasing a real estate (house or flat) is a large and relatively infrequent investment for the majority of the people. Therefore, a realistic price estimation is an extremely important task. The price of a real estate is influenced mainly by a real estate features like number of bedrooms, bathrooms, area but also some other characteristics like location and internal facilities [1, 2].

In general case the price of a real estate is determined by its owner or a real estate agent. Their estimations rely upon personal experience, interests and sometimes even emotional aspects. An objective price estimation could help both buyers and sellers to check if the proposed price is justified.

This paper proposes the model for preparing a training data set aimed creation of a regression model for predicting real estate price that includes information on attractiveness of a real estate's location as well as information on some of its additional interior facilities.

2. REGRESSION MODEL FOR REAL ESTATE'S PRICE PREDICTION

A regression model for real estate's price prediction for which the treining set is constructed is a Gradient Boosted Trees (GBT) model [3,4]. Gradient boosting works in a way to create a lower-performance

Correspondence:
Mladen Vidović

e-mail:
mvidovic@singidunum.ac.rs

DOI: 10.15308/Sinteza-2019-388-393

model first, and evaluating its output. Thereby, in gradient boosting algorithm its residual (i.e., errors) are considered instead of initial classificator's predictions. The pseudo-code of this algorithm goes as follows:

1. Using a dataset train the initial model F_0:

$$F_0(x) = y$$

2. Using residuals of an initial model train a new model $h_0(x)$:

$$h_0(x) = y - F_0(x)$$

3. Create a new model $F_1(x)$ by combining output of an initial model and a model obtained by using initial model's residuals:

$$F_1(x) = F_0(x) + h_0(x).$$

Initial models as well as the models trained using residuals are arbitrary classifiers or predictors. In the case of a *GBT (Gradient Boosted Trees)* algorithm, these models are decision trees.

3. TRAINING SET

For the purpose to establish a training set, three data sets are created here.

The first data set, the basic one, comprises data on basic real estate techical features like number of rooms, number of bathrooms, area and the similar.

The second data set contains real estate's geo-coordinates. This data set is later on used to extend the basic data set with information related to a quality of a real estate neighbourhood like vinicity of schools, restaurants, etc.

Finally, the third data set contains information both about a quality of a real estate neighbourhood, and data about its additional interior facilities.

The basic data set is created from the ads downloaded from the real estate advertising web page nekretnine.rs [5]. The Python Scrapy library [6] was used for this download. The initial URLs are retrieved from this page with a criteria to search only for flats advertised for selling in Serbian cities Belgrade and Novi Sad. Some of these elements are:

1. Price in euros (target attribute).
2. Address represented by mandatory street, city and state.
3. Number of rooms.
4. Number of bathrooms.
5. Floor.
6. Total number of floors in the building.
7. A list of corresponding photos' URLs.

There are 17 attributes in total that are described in details in [7]. The initial data set contains 95942 ads and 662396 photo URLs.

A separate data set was created of 18374 ads that contains real estate's geo-coordinates. This set was later used to collect data on real estate's location aimed at estimating its attractiveness.

The third data set that contains relevant interior facilities is determined by applying a convolution neural network to the photos attached with an ad.

4. CREATION OF A TRAINING SET

Data preprocessing

By applying explorative analysis it was observed that some ads do not have price attached, which makes them useless for training purposes, so these ads were discarded.

Next step was determination and removal of outlayers, i.e., a noise in the data. For that purpose the price and area are used considering valid only ads with declared price between 10000 and 1000000 euros and area between 10 m^2 and 300 m^2.

For the data set with data on location and detected interior facilities a correlation matrix was created that has shown that no highly correlated data exist and, hence, no removal of the data was needed.

After applying the explorative analysis results, the number of the ads left was 61942. Table I depicts an overview of the removed ads.

Textual contents of ads containing categorial and enumerative attributes are appropriately transformed. Enumerative attributes are encoded as vectors of binary values, where the value 1 denotes a presence and the value 0 denotes an absence of the corresponding attribute.

It is obvious from the Table I that the largest number of ads was removed due to missing data about the floor of the apartment, while the smallest number of removals was due to price above limit.

Table 1. Removed Data overview

Removal due to	Number of removed ads
Missing price	535
Price below limit (10000 €)	436
Price above limit (1000000 €)	24
Area below limit (10 m^2)	37
Area above limit (300 m^2)	227
Missing floor	32742

For all created data sets the minimal, maximal and average values for price and area were determined.

As shown on Figure 1, (the symbols 1, 2 and 3 denote basic, location extended, and location&internal facilities extended data respectively), prices ranges and distributions are equal in all three sets. This is what is reqired for a good training set, so our sets are acceptable for the intended purpose.

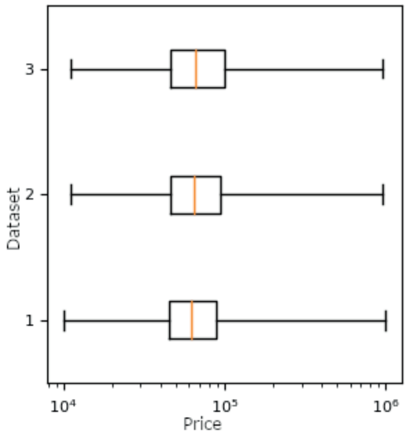

Fig. 1. Prices ranges and distributions in the data sets

Determining data about location quality

Using Overpass API [8] for OpenStreetMap [9], important objects are obtained within a given circle with a real estate as a center. Three radius values were selected, the first one from 0m do 150m, the second one from 150m to 500m, and the third one from 500m to 1000m. Within each of corresponding circles a search was performed for relevant objects and areas comprising:

1. public objects like schools, day nurseries, restaurants, hospitals, etc.;
2. Tourist objects like galleries, hotels, historic monuments;
3. Areas with a special purpose like waste disposal, industrial areas;
4. Resorts and relaxing places, sport centers, parks, paly grounds, etc.

The retrieved elements are added as attributes of a real estate to be used for the further analysis of the location quality. Thereby, only the existence of a certain object's type was considered, not the number of such objects. A proximity of the downtown was determined as an additional quality factor. The coordinates of city center are used as given by OpenStreetMap. The distance of a real estate from the city centre was computed using haversine formula for calculating the distance between two points on the sphere.

An analysis of the obtained objects in apartments' neighborhoods has shown that two larger radius are almost identical for the majority of real estate and as such of no use for the training purpose. Therefore, only the area within the smallest radius was included, i.e., only the objects within the distance of 150 meters from the real estate were considered. Also, some objects like a zoo and prison have not been detected, so they were removed. Remaining objects are added with attribute values 1 or 0, respectively indicating presence or absence of an object.

In order to avoid a huge number of attributes, a clustering of real estate based on retrieved objects in their vicinities was done. For the clustering the k-means algorithm was applied. Figure 2 shows model performance in terms of number of clusters, while the Table II depicts a clusters' overview for k=5.

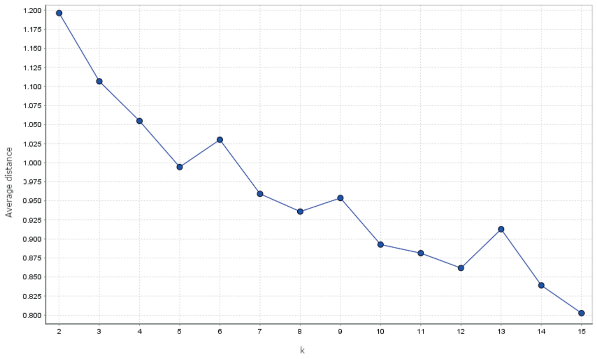

Fig. 2. Model performance in terms of clusters

Table 2.Clusters Overview For k=5

Cluster	Number of examples	Mean price	Mean area	Average price	Average area
0	1837	63860	59	77621	64.2
1	1629	85000	74	117629	78.8
2	510	71585	59.5	103373	66.6
3	2955	70560	64	93827	68.2
4	11443	60000	58	76110	63.2

In order to determine the parameter k, an iterative clustering was done with k-values varying from 2 to 15. Then the average distances of clusters' centres from the clusters' elments were considered as to observe an "elbow". From the diagram on Fig. 2 one can conclude that 5,8, and 12 are potential "elbows", i.e. optimal numbers of clusters. Finally, the number of clusters is determined empirically.

Detecting interior facilities from the photos

A set of 7998 annotated photos is used for training a convolution neural network with 25% of photos marked off for testing. Some labels are removed because no real estate with those labels were present in the data set.

The Tensorflow Object Detection API [10] was used for dealing with network. In order to reduce the training time, the network was used that was pre-trained on the COCO (Common Objects in COntext) data set containing objects to be detected on the photos [11]. Afterwards, a fine tuning of the network was done, i.e. the weights of the shallow layers were retained for basic features extraction, and the deeper layers are than additionally trained using photos from ads.

The pre-trained Faster R-CNN ResNet101 [12] with 101 layers was used, characterized with a good performance on the COCO data set and relatively short training time per photo.

After each convolution layer a batch normalization is performed.

The Rectified Linear Unit is used as an activation function of convolution layers.

The network also has two pooling layers for dimension reduction, one on the input where the max pooling is done with a kernel size 3 and step 2, and one on the output where average pooling is performed with kernel size 7 and step1.

One fully connected layer with a softmax activation function is on the network output.

The network input was fed with manually annotated photos form ads for the additional network training.

In order to reduce input layer, actually to reduce a training time and hardware requirements, the images are scaled to have maximal height and width of 800 pixels and minimal height and width of 600 pixels. Prior to training, an augmentation of data set was performed aimed at creating a more robust model and avoiding network over-fitting. The augmentation method that were applied are:

- Image horizontal flip, and
- Image illumination change.

The training was done in 50000 epochs.

An example of the detected interior facilities is shown on Figure 3.

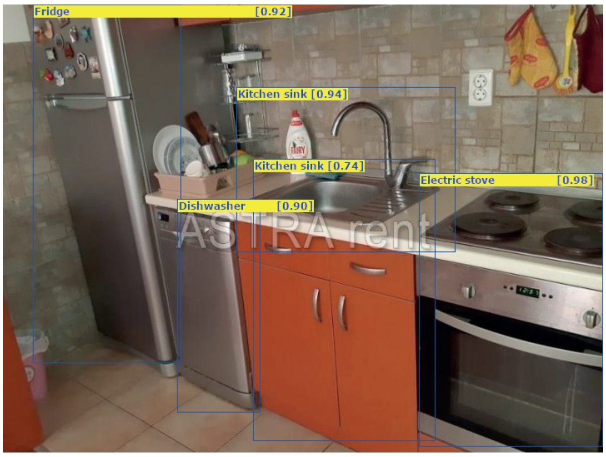

Fig. 3. An example of the detected interior facilities

Once the network was trained, it is used for detecting objects on the photos from 9862 ads having both geodata and images. Images from these ads were not the part of the network training set and, consequently, they were not annotated.

The performance evaluation for the network was carried out using the data set previously marked off for a testing and not used for the training. Two metrics are used for evaluation. The first one is AP (average precision) that applies to a single class, while the second one is MAP (*Mean Average Precision*) which applies to all classes and represents an average of APs. The results for the achieved AP values are presented in [5], while the MAP value for this data set is 0.594.

5. RESULTS OF THE DATA USE FOR TRAINING A GBT MODEL

Model training and evaluation are done for the basic data set, data set with a neighborhood attractiveness estimation, and the data set with a neighborhood attractiveness estimation and internal facilities detected from the photos.

Each data set was split into a training set with 80% of elements and a test set with 20% of elements. Then, additional 20% of elements are pulled out to form a validations set.

Model is trained on the training set and its parameters are then optimized on the validation set. Optimization was iterative by changing the values of the relevant parameters and comparing the model's performance on the validation set. The metrics used for performance evaluation was R^2. The GBT ensemble is optimized by the depth of the regression three and number of estimators.

For the basic data set the optimal tree depth was 12 and the number of estimators was 70. For the data set with geo-coordinates but no interior facilities the optimal tree depth was 9 and the number of estimators was 80 both before and after adding a cluster label. Finally, for the data set containing both geo-coordinates and interior facilities with only basic data about real estate, optimal three depth is 12 and the number of estimators is 70. After adding data on location quality the optimal three depth of 11 and number of estimators of 60 were obtained. With all available data, optimal depth is 12 and optimal number of estimators is 60. Table IV depicts the achieved performance for various data sets.

Table 3. GBT Model Performance For Various Data Sets

Data set	R^2 achieved on validation set
Basic data	0.77
Geo-coordinates, no clusters	0.767
Geo-coordinates with clusters	0.856
Geo-coordinates and images, basic data	0.768
Geo-coordinates and images, with clusters, no interior facilities	0.839
Geo-coordinates and images, with all available data	0.816

6. CONCLUSIONS

In this paper a model for constructing training set is proposed for predicting an objective price of real estate from Internet ads relying upon its features. Real estate features include, along with its basic technical characteristics, the attractiveness of a real estate's location and interior facilities extracted from the photos accompanying the advertisements.

A real estate's location attractiveness determination is based on relevant objects located within a circle with 150 meters radius. Those objects are determined using the Overpass API for OpenStreetMap.

For detection interior facilities of a real estate, a convolution neural network was used. The MAP (*Mean Average Precision*) achieved by the network was 0.594.

There are several directions that could be taken to improve the quality of the training set.

Firstly, a larger number of the annotated images could improve the network's performance, while the feature interior facilities could also be improved by creating a unified estimate using, for example, a clustering technique.

Real estate's location could be used to derive some additional attributes lake prices of the surrounding real estate, which according to some previous research proved to be an important factor of the price.

Finally, authors' impression is that a free text describing the real estate contains substantial information about its basic features, which indicates that text mining techniques could be useful for improvement of the quality of data set.

REFERENCES

[1] J. Ottensmann, S. Payton, J. Man, "Urban Location and Housing Prices within a Hedonic Model", Journal of Regional Analysis and Policy, vol. 38, no. 1, pp. 1-17, 2008.

[2] F. Kong, H. Yin, N. Nakagoshi, "Using GIS and Landscape Metrics in the Hedonic Price Modeling of the Amenity Value of Urban Green Space: A Case Study in Jinan City, China", Landscape and Urban Planning, vol. 79, no. 3, pp. 240–252, 2007.

[3] Y. Freund and R. E Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", Journal of computer and system sciences, vol. 55, no. 1, pp. 119–139, 1997.

[4] J. H. Friedman, "Greedy Function Approximation: a Gradient Boosting Machine", Annals of statistics, vol. 29, no. 5, pp 1189–1232, 2001.

[5] Nekretnine. (2018). [Online]. Available: https://www.nekretnine.rs.

[6] Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. (2018). Scrapinghub, Ltd. [Online]. Available: https://github.com/scrapy/scrapy

[7] M. Vidović, "Prediction of a Real Estate Price Based on Data from Avertisings", (in Serbian), M.S. Thesis, Faculty of Technical Sciences , University of Novi Sad, Serbia, 2018.

[8] Overpass-API. (2019), OpenStreetMap Foundation. [Online]. Available: https://wiki.openstreetmap.org/wiki/ Overpass_AP.

[9] OpenStreetMap. (2019), OpenStreetMap Foundation. [Online]. Available: https://www.openstreetmap.org.

[10] J. Huang et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI , pp. 3296–3297..

[11] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, pp. 740–755.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778.