APPLICATION OF UNSTRUCTURED TEXT BASED FEATURES IN PREDICTION OF REAL ESTATE PRICES: A COMPARATIVE STUDY

Diana Vranešević, Đorđe Nedeljković, Miloš Kovačević

Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia
e-mail: dvranesevic@grf.bg.ac.rs, ndjordje@grf.bg.ac.rs, milos@grf.bg.ac.rs

Abstract:

This study demonstrates the potential of application of unstructured textual data for predicting real estate prices and compares different protocols for extracting features from textual data. Performance of the different models for price prediction was evaluated on data set of real estate listings, which included numerical and categorical features, as well as text descriptions. The experiments showed that adding features extracted from both the translated description text, as well as noun chunks from it, resulted in the highest R^2 score of 0.768, representing an improvement over the R^2 score of 0.71 for the baseline model without text-based features. The findings from this study indicate how the performance of real estate price prediction models can be improved by utilizing text-based features, in turn benefiting property market stakeholders in making informed decisions and evaluating competitive pricing strategies.

Key words: real estate price prediction, ridge regression, NLP, text feature extraction

1. Introduction

Real estate prices are a crucial factor in many financial and investment decisions. In recent years, the availability of copious amounts of data and advances in machine learning techniques have opened new possibilities for progress in predicting real estate prices. The accurate price prediction is a challenging task, due to the many factors that can influence it, such as location, size, amenities, etc. Improved price prediction performance can specifically benefit buyers, sellers, and investors. Investors and sellers can use it to evaluate either the most promising areas of a city for new property development or to determine the most competitive pricing strategy for their properties. Potential buyers can also benefit, by having nonsubjective assessment of real estate price for their desired parameter values and property description, allowing them to make an informed decision to suit their needs while saving significant amount of time and effort.

Methods for real estate price prediction are mostly AI based [1], with various regression methods being among the most used [2]. In addition to numerical and categorical features mostly used by regression methods, regression methods for price prediction can benefit from introduction of unstructured textual data [3]. This study compares and analyzes performance of different models for real estate price prediction with included features extracted from unstructured textual data.

2. Data Collection

In this study publicly available data from the online real estate listing site[1] were collected. The scope of data collection was apartments in urban municipalities in Belgrade, Serbia, listed on February 1st, 2023. Figure 1 shows locations of apartments on the Belgrade city map, with the price per m^2 in EUR. The total size of the dataset was 1800 apartments.

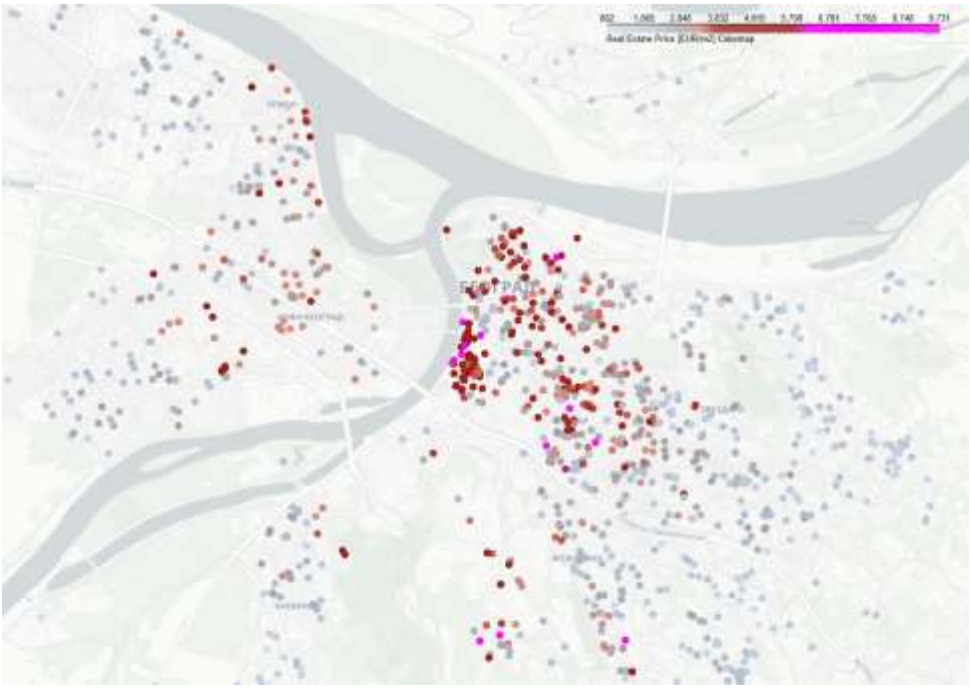

Fig. 1. Listed real estates on February 1st, 2023, with prices ranging from 882 to 9731 EUR per m^2

Each extracted record had following feature:

- Categorical: municipality, neighborhood, street, structure (ranging from studio to 5+ rooms), floor, total floors, building state (existing, new), VAT return (yes, no)
- Numerical: size in m^2; average EUR per m^2 price for the neighborhood
- Textual: description

3. Baseline Model and Evaluation Metric

Prior to modeling, data was preprocessed by normalizing the numerical features and hot encoding the categorical features. The baseline model used only the numerical and categorical features extracted from the real estate listings (NCF model).

RidgeCV regression method was used for real estate price prediction. The method was implemented through scikit-learn library, with default hyperparameters[2]. To evaluate the performance of models, a cross-validation approach was used, with a Coefficient of determination (R^2) score, provided by RidgeCV. The R^2 score was chosen as the evaluation metric due to its suitability for regression problems and ease of interpretation. A five-fold cross-validation scheme

[1] https://www.halooglasi.com/nekretnine
[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html

was used, with a mean score from 100 iterations to ensure robustness of the results.

The baseline model achieved an R^2 score of 0.71, providing a starting point for comparison with the models that used additional textual features.

4. Textual Feature Extraction

The potential of unstructured textual data for real estate price prediction was explored through four different protocols for extracting features from the textual descriptions.

Scikit-learn's TfidfVectorizer[3] class was used to extract features from the raw textual data. In the first protocol, TfidfVectorizer was applied on the original text in Serbian language (NCF_TF_O model). In the second protocol, the same procedure was executed on the descriptions automatically translated to English (NCF_TF_T model).

The last two protocols utilized the noun chunks extracted from the translated text. In the context of the natural language processing, a noun chunk is a contiguous sequence of words in a sentence that contains a noun and any words that modify the noun. In this research, noun chunks extraction was performed using spacy[4] library.

The third model was generated by applying TfidfVectorizer solely on noun chunks (NCF_TF_NC model). Finally, the last protocol combined both the features extracted from the translated text with the noun chunks features (NCF_TF_TNC model). All the models used in the research are shown in Figure 2.

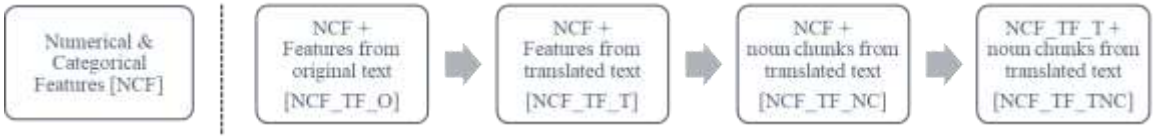

Fig. 2. Descriptions of baseline (with only numerical and categorical features) and expanded models (with included features extracted from the real estate description)

5. Results and Analysis

The performance of the models was evaluated using the R^2 score obtained from cross-validation. Experiments indicate that adding features extracted from the translated text or combining features from the translated text and noun chunks, led to the highest R^2 scores of 0.759 and 0.768, respectively (Table 1). These scores represent an improvement over the baseline R^2 score of 0.71.

The model NCF_TF_NC that used only features from the noun chunks performed worse than other models with text-based features, suggesting that the noun chunks alone are not as informative as raw textual description for the task of real estate price prediction. However, highest performance of the NCF_TF_TNC model, which used features from both the translated text and noun chunks, indicates that noun chunks carry unique informative meaning, which can further augment the informativeness of the singular words used as the model features.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[4] https://spacy.io/usage/linguistic-features#noun-chunks

Protocol	NCF	NCF_TF_O	NCF_TF_T	NCF_TF_NC	NCF_TF_TNC
R^2 score	0.71	0.745	0.759	0.742	**0.768**

Table 1. R^2 scores for baseline and four models with text-based features for real estate price prediction

6. Conclusion

In this study, the use of textual features for predicting real estate prices was explored. Overall, experiments demonstrate that text-based features extracted from real estate listings contain valuable information, relevant for predicting real estate prices. Furthermore, the results suggest that translating the text and extracting noun chunks can be an effective approach for improving prediction accuracy. Both the singular words and noun chunks contribut individually to informativeness of the model, indicated by the highest performing model which combined features from both the aforementioned text representations.

Overall, the advancements towards creating robust and high-performance real estate price prediction models have the potential to positively impact the stakeholder's actions in property market. More reliable price prediction models can increase the competitive edge for the investors, sellers, and potential buyers, by allowing them to make more strategic and informed decisions in a highly competitive and dynamic industry.

References

[1] Wang D, Li VJ. Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. *Sustainability*. 2019; 11(24):7006. https://doi.org/10.3390/su11247006

[2] Mohd, T., Jamil, N.S., Johari, N., Abdullah, L., Masrom, S. (2020). An Overview of Real Estate Modelling Techniques for House Price Prediction. In: Kaur, N., Ahmad, M. (eds) Charting a Sustainable Future of ASEAN in Business and Social Sciences. Springer, Singapore. https://doi.org/10.1007/978-981-15-3859-9_28

[3] Xuerong Li, Wei Shang, Shouyang Wang, Text-based crude oil price forecasting: A deep learning approach, International Journal of Forecasting, Volume 35, Issue 4, 2019, Pages 1548-1560, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2018.07.006.