Machine learning model for house price predicting based on natural language text data analysis

Sergey Bushuyev[1,*,†], Denis Bushuiev[1,†], Dmitriy Kravtsov[2,†] , Nikolay Poletaev[2,†] and Mykola Malaksiano[2,†]

[1] *Kyiv National University of Construction and Architecture, 31, Povitroflotskyi avenue, Kyiv, Ukraine*
[2] *Odessa National Maritime University, 34, Mechnikov street, Odesa, Ukraine*

Abstract

A machine learning model is proposed to enhance the accuracy of predicting rental prices of real estate based on the analysis of textual data processed using the Term Frequency-Inverse Document Frequency method. This work explores the enhancement of predictive model effectiveness by integrating detailed textual information into a basic dataset of numerical features. Utilizing the Light Gradient Boosting Machine method for regression analysis, it was found that adding textual features significantly reduces the Mean Squared Error, demonstrating an improvement in the predictive capabilities of the model. The value of this work is defined by a comprehensive approach to analysis, which combines textual and numerical data for a deep interpretation of the impact of features on real estate rental pricing.

Keywords

machine learning, neural networks, natural language text data analysis, predicting model, house price, regression analysis, light gradient boosting machine

1. Introduction

The real estate market attracts significant investment volumes and is a significant part of the economy of every state. In this context, new methods aimed at studying pricing trends and optimal management of investment projects in the field of real estate operations are actively developed and implemented. In recent years, a trend in real estate pricing forecasts has emerged towards integrating textual data with basic numerical features, reflecting an aspiration for a more comprehensive and multifactorial analysis of price trends. This direction emphasizes the importance of information contained in textual descriptions of real estate properties and explores how this data can be used to improve the accuracy of estimation models. Research [1] in this area demonstrates how textual descriptions of real

MoMLeT-2024: 6th International Workshop on Modern Machine Learning Technologies, May, 31 - June, 1, 2024, Lviv-Shatsk, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

sbushuyev@ukr.net (S. Bushuyev); bushuyevd@gmail.com (D. Bushuiev); dmkravtsov@gmail.com (D. Kravtsov); poletaev@ukr.net (N. Poletaev); malax@ukr.net (M. Malaksiano)

0000-0002-7815-8129 (S. Bushuyev); 0009-0002-6477-0517 (D. Bushuiev); 0009-0005-0305-544X (D. Kravtsov); 0000-0002-1340-582X (N. Poletaev); 0000-0002-4075-5112 (M. Malaksiano)

estate can influence valuation, especially when combined with non-textual data such as size or location of the property. The use of advanced text processing technologies, including word vectorization methods such as TF-IDF, Word2Vec, and BERT, has significantly improved the quality of forecasts. The gradient boosting algorithm, which combines textual and numerical data, shows impressive results in prediction accuracy. Similar conclusions were drawn in another study [2], which emphasizes that including textual information in real estate valuation models can significantly improve their efficiency. This approach opens new opportunities for identifying unique characteristics of properties that are usually not considered in traditional valuation methods. Problems concerned with sentence parsing for determining keywords, automated identification of metaphoric meaning, ontology learning, and knowledge evaluation were studied in [3, 4, 5].

A separate study [6] represents an innovative attempt to combine visual and textual data for real estate valuation. This approach demonstrates that supplementing textual information with visual elements can significantly improve the accuracy of evaluations, providing a more comprehensive understanding of the property's value characteristics.

The analysis of geolocational textual data also plays an important role, as shown by research [7]. Here, the use of geotagged Wikipedia articles has allowed for the enhancement and expansion of GIS methods, providing a more detailed understanding of local features that may affect real estate prices. Finally, research [8] highlights the importance of sentiment analysis and preliminary text processing in the context of assessing real estate market trends. Such an approach allows not only to collect but also to deeply analyze textual data, obtaining valuable insights about market sentiments and trends.

These studies demonstrate the significant potential of textual data in improving real estate pricing prediction methods, opening new perspectives for research and practical application.

Alongside the development and implementation of modern information technologies and artificial intelligence methods, there is also currently a significant practical and scientific interest in the corresponding improvement of methods for managing investment projects. Works [9, 10] address the issues of managing infrastructure projects in conditions of uncertainty. Mechanisms for Goal Setting, Synergetic Effects, and Risk Management of Concession Projects were studied in [11, 12, 13]. Articles [14, 15] propose a dynamic model for projects portfolio structure provided a resistance of information entropy and develop Entropy paradigm of project-oriented organizations management. Issues of eco-oriented project management and the formation of balanced development trajectories were studied in [16, 17]. Problems of modeling aging processes and depreciation of infrastructure facilities and creating appropriate information systems for decision-making support were explored in [18, 19]. SMART intelligence models for managing innovation projects were studied in [20] based on groups of interrelated competencies. Paper [21] discusses the genetic approach application and creation of the project genetic model.

2. Machine learning model and data interpretation considering textual data

2.1. Methodology for generating new features based on textual data

During the study, a detailed analysis and transformation of the data in a set consisting of 9,260 objects and 10 features, including the target feature "Price", which describes properties in Houston, USA, were conducted (Table 1). The data were obtained from the Redfin real estate company's website [22]. Each feature was carefully examined and processed to eliminate incorrect, anomalous, or missing values, in order to maintain data interpretability.

Table 1
Description of basic features

Feature	Description	Data Type	Units
Latitude	Geographic latitude of the property	float	Degrees
Longitude	Geographic longitude of the property	float	Degrees
YearBuilt	The year in which the property was constructed	int	Year
Beds	Number of bedrooms in the property	int	Count
Baths	Number of bathrooms in the property	int	Count
BuildingSize	The size of the property's building	float	Square Feet
LotSize	The size of the property's land/lot	float	Square Feet
PostalCode	Postal code where the property is located	string	-
Description	Textual description detailing the property	string	-
Price	The rental price of the property	float64	U.S. Dollars

As an example of a descriptive feature, consider the following description of one of the real estate properties (Feature Description):
"These apartments feature an open concept living/dining room, and a kitchen with granite countertops, subway tile backsplash, and stainless-steel appliances with a dishwasher. Features include central A/C, central heat, and ceiling fans. The bedroom(s) have carpet flooring, and the rest of the home has hardwood floors. Select units offer nine-foot ceilings, built-in desks, shelving, and a private balcony. The community provides elevators, wheelchair access, a saltwater rooftop pool with sundeck, rooftop tennis court, and more."

To prepare data effectively for analysis and subsequent modeling, a meticulous transformation and encoding of the textual descriptions contained in the dataset's Description feature were carried out. The process of processing textual data included the following key stages:

- Text Tokenization: this step involved breaking down textual descriptions into individual words or terms. Tokenization allowed for efficient text processing by converting it into a sequence of lexemes or tokens. This is necessary for subsequent text processing, as computer models work better with individual words rather than entire sentences or paragraphs.

- Stop-word Removal: stop words are words that do not carry significant semantic load in the context of analysis (e.g., prepositions, conjunctions, and frequently used words like "and", "in", "on"). Removing them helps to reduce noise in the data and focus on more significant words for analysis.
- Word Lemmatization: lemmatization is the process of converting words to their base, or lemmatized, form. This means that words in different forms (e.g., plural, various tenses, etc.) are transformed into a single basic form.

The combination of these methods improved the quality of textual data, which provided a deeper and more accurate analysis for subsequent machine learning and real estate rental price prediction.

To convert textual descriptions of real estate properties into a numerical format suitable for machine learning algorithms, the TF-IDF vectorization method was used [23]. This approach allowed assigning a specific weight to each word or phrase in the text, reflecting their importance in the context of the entire dataset. TF-IDF increases the significance of words that are unique or important to a specific document but are not frequently encountered in many other documents, making this measure particularly useful for text analysis and information retrieval.

2.2. Algorithm for predicting real estate prices

The algorithm for predicting the rental cost of real estate includes the following stages:

1. Using RFECV with Decision Tree Algorithm: the Recursive Feature Elimination with Cross-Validation (RFECV) [24] was utilized with a Decision Tree (DT) as the base classifier. RFECV is a feature selection technique that recursively removes the least important features and evaluates the model at each stage using cross-validation. This approach helps determine the optimal number of features and their combination to achieve the best model performance.
2. Model Hyperparameter Optimization Using Optuna: for optimizing the model's hyperparameters, Optuna [25], an automated hyperparameter optimization tool, was chosen. Optuna outperforms methods such as RandomizedSearch and GridSearch by more efficiently searching the parameter space and allowing for more complex strategies like pruning trials that do not appear promising. This reduces the time required to find optimal parameters and increases the likelihood of discovering the best parameter combination for the model. The authors employed Optuna along with the LightGBM model (Light Gradient Boosting Machine), selected for its high performance, efficiency, and support for handling a large number of features [26], which was particularly important given the enriched dataset context.

These steps resulted in an enhanced prediction model for real estate rental pricing, with the most significant features identified and the model parameters optimally tuned.

2.3. Algorithm for interpretation of results

The interpretation of results involves the following steps:

1. Using SHAP TreeExplainer: the authors utilized SHAP (SHapley Additive exPlanations) [27] with TreeExplainer for the LightGBM model. SHAP TreeExplainer provides a detailed interpretation of model predictions by calculating the contribution of each feature to each specific forecast. This is done using Shapley values - a concept from game theory that distributes the "payoff" (or influence) among all features. The advantage of SHAP lies in its ability to explain the behavior of the model at the level of individual predictions, making the results more understandable and interpretable.
2. Application of Partial Dependence Display: the Partial Dependence Display method [28] was also used to visualize dependencies between the most important features and the target variable. This method allows isolating and examining the influence of a single feature, while averaging the effects of all other features. This makes its interpretation more straightforward and clearer compared to methods that consider the combined influence of all features.

This approach is most useful when there is already a general understanding of the importance and influence of features in the model, as it provides an opportunity to deepen the understanding of how a specific feature affects the forecast and explore non-trivial forms of dependencies (e.g., nonlinear). Therefore, its use is appropriate in the final stages of analysis when a detailed examination of the influence of specific features is needed.

By combining these methods, the authors gained a deep understanding of how the model works, identified key factors affecting the cost of renting real estate, and improved the interpretability of forecasting results.

Thus, the proposed model for assessing the values of real estate operations projects, based on natural language text data, is depicted in the following scheme (Figure 1).

3. Results and Discussion

The application of the TF-IDF algorithm to text processing resulted in the creation of 845 new unique textual features, each assessing the frequency of inclusion of a particular word within the context of the entire text corpus, considering its importance within an individual document and its frequency across all documents in the dataset.

Using the RFECV method on the expanded feature set allowed for the reduction of these features from 845 to the 38 most significant ones, including both basic features (Table 1) and new textual features such as 'stainless', 'elevator', 'laundry', 'concierge', 'pool', among others, as well as 'sentiment_score', which reflects the overall emotional tone of the real estate descriptions.

By applying cross-validation on five different data subsets, key performance indicators of the model were calculated, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and Median Absolute Percentage Error (MDAPE). After 400 iterations using the Optuna algorithm, the best combination of parameters that minimized MSE was found. This process ensured maximum accuracy and efficiency of the predictions for the LightGBM model.

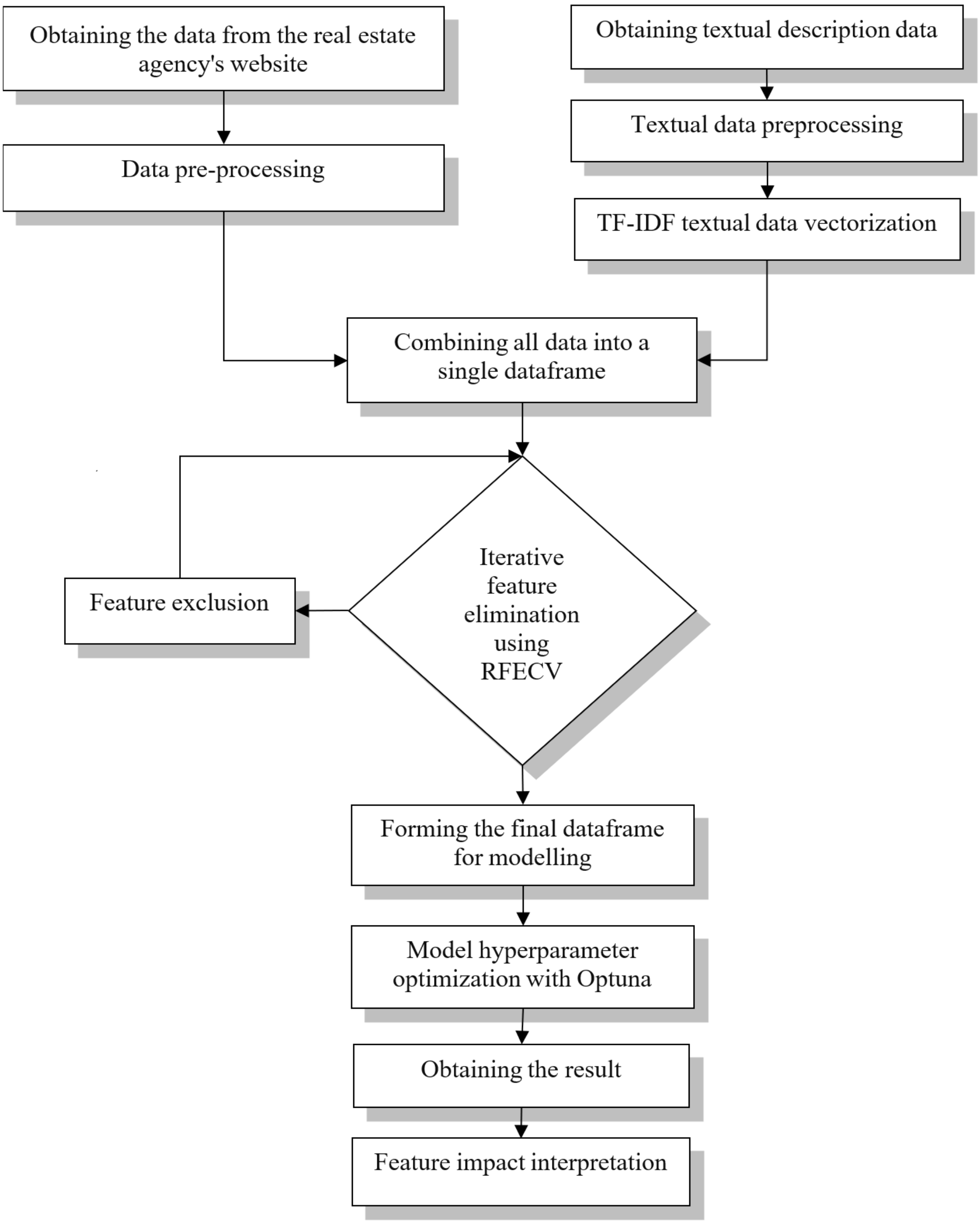
Obtaining the data from the real estate agency's website
Data pre-processing
Obtaining textual description data
Textual data preprocessing
TF-IDF textual data vectorization
Combining all data into a single dataframe
Iterative feature elimination using RFECV
Feature exclusion
Forming the final dataframe for modelling
Model hyperparameter optimization with Optuna
Obtaining the result
Feature impact interpretation

Figure 1: Scheme of the model for assessing real estate operations projects values based on natural language text data.

From Table 2, it's evident that the method of enriching data with textual features has improved the model's metrics across all key measures. The Mean Squared Error (MSE), which assesses the square of errors, showed an improvement of 13.4%, indicating a significant reduction in large errors. The Root Mean Squared Error (RMSE), which brings errors to their original units, also decreased by 6.95%. An increase in the coefficient of determination (R^2) by 1.96% indicates that the model has become better at explaining the variability of the data. The reduction in Median Absolute Percentage Error (MDAPE) by 12.46% reflects the improved accuracy of the model, particularly due to its lesser sensitivity to extreme values compared to MSE and RMSE.

Table 2
Model efficiency depending on the selected metric

Parameter	Before Textual Feature Enrichment	After Textual Feature Enrichment	% Improvement
MSE	139998	121223	13.4
RMSE	374	348	6.95
R^2	0.818	0.834	1.96
MDAPE	7.830	6.854	12.46

In the SHAP value plot (Figure 2), we see the distribution of the contribution of each feature to the predicted rental cost. Positive SHAP values (blue dots) indicate that an increase in the feature leads to an increase in rental cost, while negative SHAP values (red dots) suggest the opposite trend:

- 'BuildingSize' and 'Baths' have a significant number of positive SHAP values, indicating that a larger building size and more bathrooms increase the rental cost. This suggests that these attributes are highly valued in the rental market, possibly due to the added comfort and functionality they provide to tenants.
- 'PostalCode', 'Beds', 'Latitude', and 'Longitude' also influence the cost, highlighting the importance of location and the number of bedrooms in assessing rental value. These factors underscore the role of geographical positioning and sleeping capacity in determining rental prices, reflecting both the desirability of the area and the practical needs of potential renters.
- The contribution of 'YearBuilt' is relatively smaller but still noticeable. The modernity or newness of a building may play a role in its value, with newer properties potentially offering more up-to-date amenities, better energy efficiency, and fewer maintenance issues, which are appealing traits for renters.
- Features related to amenities, such as 'elevator' and 'stainless' (possibly referring to stainless steel appliances or fixtures), have lower SHAP values but contribute to the overall cost. While these might not be as crucial as size or location, they still add value by enhancing the lifestyle and convenience for residents, thereby affecting rental prices to a certain extent.

These insights from the SHAP value analysis provide a detailed view of what features drive rental costs and how significant each one is in the model. This level of interpretation helps in refining marketing strategies, property improvements, and even future developments to align with what truly impacts rental pricing according to market data.

The dispersion of points along the SHAP value axis can reflect the varying influence of a feature in different contexts; for example, the size of the building ('BuildingSize') may have a greater impact on cost in one location and less in another. The vertical bar indicating zero influence shows that features closer to this line have a lesser impact on the model's prediction.

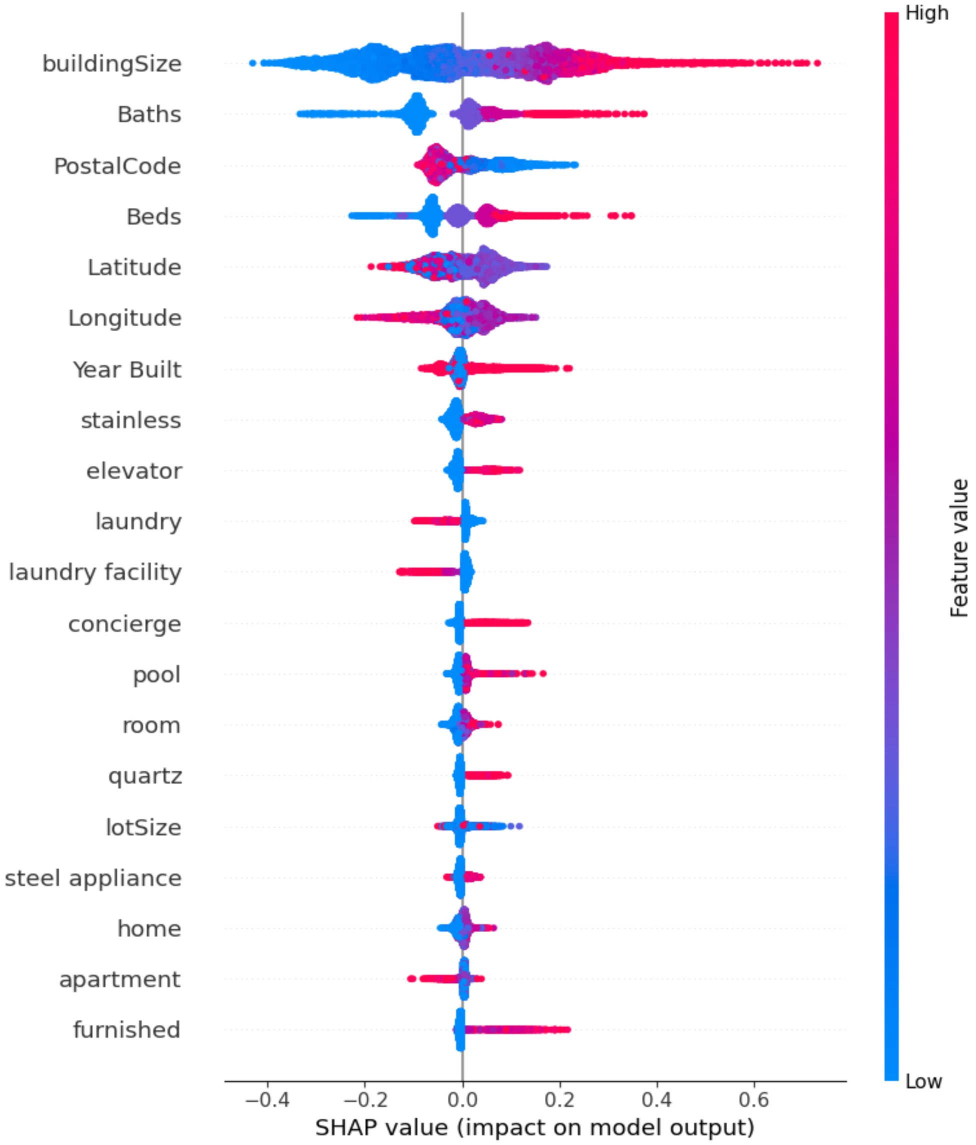

Figure 2: SHAP value (TOP-20 features).

From the partial dependence plot (Figure 3), we can judge how the size of the feature (in this case, 'BuildingSize') affects the target variable. It can confidently be stated that as the building size ('BuildingSize') increases, the value of the cost also rises. However, after reaching a certain point, approximately at a 'BuildingSize' value of about 1600, this trend stabilizes and even slightly decreases, which may indicate the presence of a threshold building size beyond which further increases do not significantly impact the rental cost.

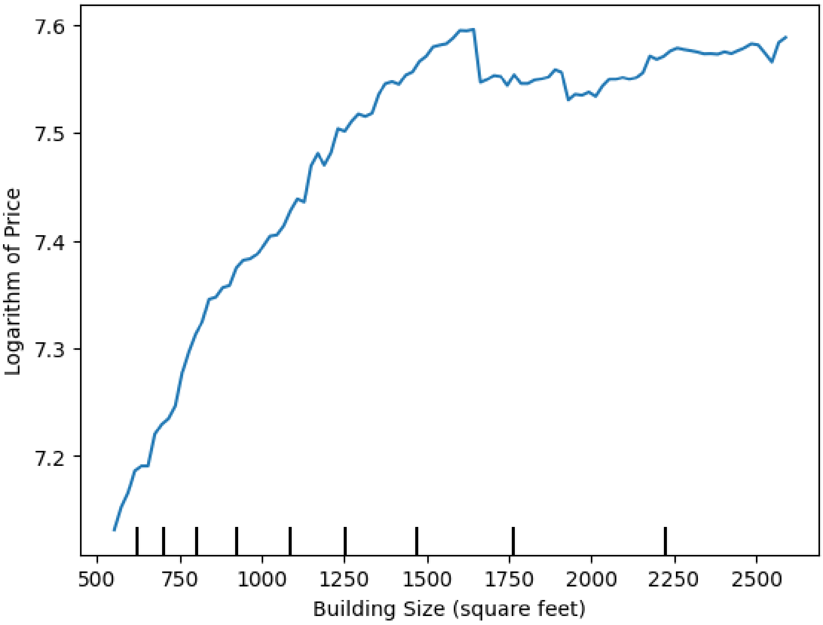

Figure 3: Partial dependence (feature - 'buildingSize').

Overall, the partial dependence method is useful for understanding the monotonic relationship between a feature and the target variable. In this case, it allows focusing on the influence of one particular feature and provides a simpler and clearer interpretation of its effect compared to other methods that might consider complex interactions between features.

The partial dependence analysis of four key features (Figure 4), including 'Latitude' and three textual features with TF-IDF weights ('elevator', 'laundry', and 'pool'), revealed the following trends affecting the logarithm of real estate prices:

- 'Latitude': Partial dependence shows a peak in a specific range of latitude, indicating the presence of a particular location where real estate is valued higher. This could reflect market preferences or the presence of important infrastructure facilities in that area.
- 'Elevator': An increase in the TF-IDF weight of this word in descriptions generally accompanies an increase in cost, emphasizing the value of having an elevator as an important amenity, especially in high-rise buildings.
- 'Laundry': Initially, there is a drop in price with an increase in the TF-IDF weight of the word 'laundry', which could be associated with less demanded or more accessible properties where laundry services are a standard amenity.
- 'Pool': The weight of this word in descriptions consistently associates with a higher price, confirming that the presence of a pool is perceived as a feature of luxurious and desirable real estate.

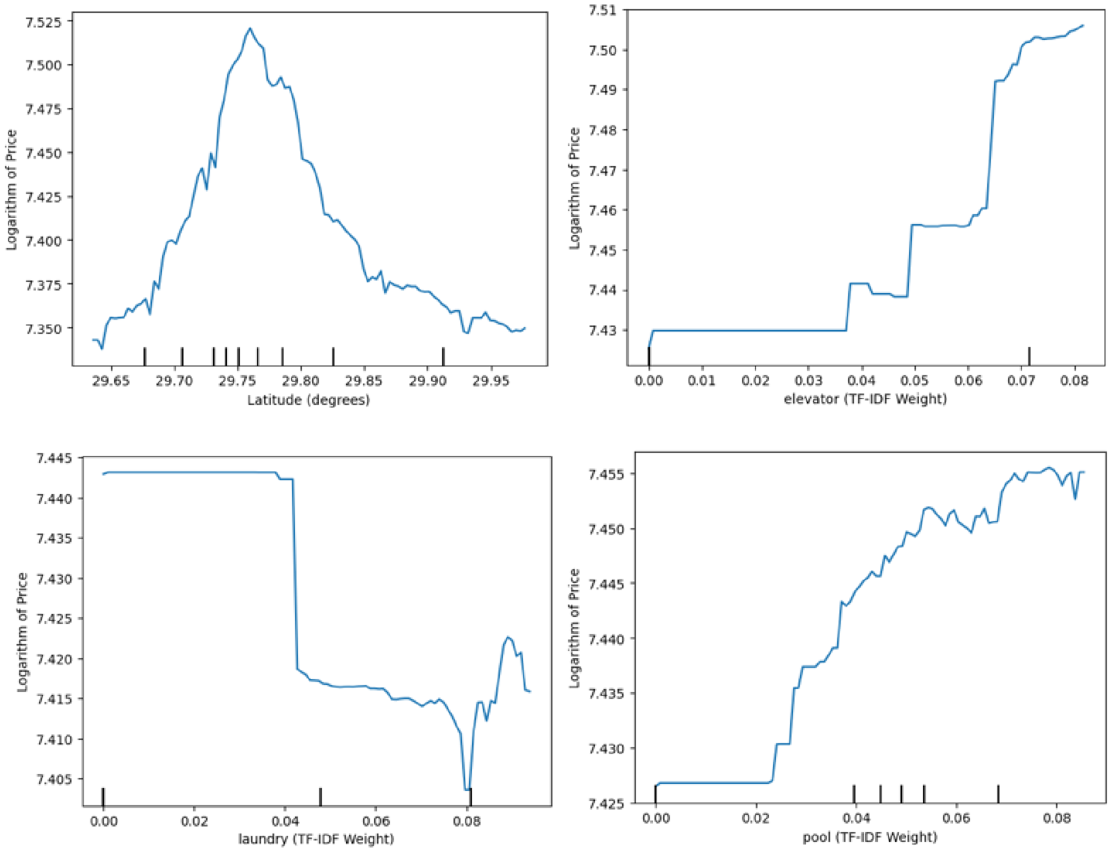

Figure 4: Partial dependence ('Latitude', 'elevator', 'laundry', 'pool features')

These four features collectively show how location and key amenities mentioned in descriptions can influence the perception of value and attractiveness of real estate. The presence of an elevator and a pool has a positive impact on the cost, while the perception of laundry services may depend on the market context. Latitude acts as a geographical indicator that may reflect proximity to important areas or urban infrastructure centers.

4. Conclusions

The proposed machine learning model, based on the analysis of textual data, has proven effective in predicting the rental cost of real estate. It was found that adding textual features reduces the Mean Squared Error (MSE) by 13.4%, demonstrating an improvement in the model's predictive capabilities for the dataset studied.

In this research, comprehensive work was conducted on analyzing the impact of various factors on the cost of renting real estate:

- Data Preparation: the data were meticulously prepared for analysis, which included transforming and encoding textual descriptions using the TF-IDF method, allowing the text to be converted into numerical features for further use in machine learning.

- Modeling and Optimization: predictive models were built and optimized using the LightGBM algorithm, which included tuning hyperparameters with the Optuna library to achieve better predictive quality.
- Feature Interpretation: various methods were applied to interpret the influence of features on the target variable, including partial dependence and SHAP values, which allowed for a deeper understanding of the contribution of individual real estate characteristics to the predicted cost.
- Partial Dependence Analysis: partial dependence plots for key features such as latitude, the presence of elevators, laundry facilities, and pools were visualized and analyzed, revealing their varying influence on the rental price.
- General Conclusions: the results of the analysis underscored the significance of location, housing size, and the presence of certain amenities mentioned in real estate descriptions in determining rental costs. This includes the positive impact of factors such as elevators and pools and the opposite influence of elements such as the presence of laundry services.

The work demonstrated how thorough data analysis, modeling, and interpretation can help create more accurate and informative predictive models in real estate. The findings can be used to improve business strategies, operational real estate management, and the development of effective pricing algorithms. It should be noted that descriptive texts, written by different people for various properties, introduce elements of subjectivity and heterogeneity into the prediction model. Differences in individual styles, levels of detail, and preferences in descriptions can complicate the analysis of textual data, as they create variability that does not always correlate with the actual characteristics and value of real estate. Such variability may require the application of advanced text processing methods to correctly highlight and standardize informative features for use in analytical models.

The significance of this research lies in the combination of methods for integrating textual data and their subsequent interpretability alongside basic numerical features, which enhances the accuracy of valuations in real estate. This allows not only to account for additional amenities and features of properties but also to explain their role in shaping market value, finding application in various fields, from improving valuation algorithms to developing asset management strategies.

References

[1] H. Zhang, Y. Li, P. Branco, P. Branco, Describe the house and I will tell you the price: House price prediction with textual description data. Natural Language Engineering (2023). doi:10.1017/S1351324923000360.

[2] L. F. Bittencourt, O. Parraga, D. D. Ruiz, I. H. Manssour, S. R. Musse, R. C. Barros, Leveraging Textual Descriptions for House Price Valuation. In Intelligent Systems (2022) 355–369. doi:10.1007/978-3-031-21686-2_25.

[3] D. Dosyn, V. Lytvyn, V. Kovalevych, O. Oborska, R. Holoshchuk, Knowledge discovery as planning development in knowledgebase framework. 13th International Conference on

Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Lviv, Ukraine, 2016, pp. 449–451. doi:10.1109/TCSET.2016.7452085.

[4] O. Levchenko, N. Romanyshyn, D. Dosyn, Method of automated identification of metaphoric meaning in adjective + noun word combinations (based on the Ukrainian language). Workshop of the 8th International Conference on "Mathematics. Information Technologies. Education": Modern Machine Learning Technologies and Data Science, MoMLeT and DS, CEUR Workshop Proceedings, 2386, 2019. URL: https://www.scopus.com/record/display.uri?eid=2-s2.0-85068030426&origin=resultslist.

[5] V. Lytvyn, V. Vysotska, D. Dosyn, R. Holoschuk, Z. Rybchak, Application of sentence parsing for determining keywords in Ukrainian texts. 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2017, pp. 326–331. doi:10.1109/STC-CSIT.2017.8098797.

[6] E. Ahmed, M. N. Moustafa, House Price Estimation from Visual and Textual Features. In NCTA, 8th International Conference on Neural Computation Theory and Applications, 2016. doi:10.5220/0006040700620068.

[7] T. Heuwinkel, J.-P. Kucklick, O. Müller, Using Geolocated Text to Quantify Location in Real Estate Appraisal. In 55th Hawaii International Conference on System Sciences (HICSS-55), 2022. doi:10.24251/HICSS.2022.700.

[8] N. Sinyak, T. Singh, J. M. Kumari, V. Kozlovskiy, Predicting real estate market trends and value using pre-processing and sentiment text mining analysis. Real Estate Economics Management, 2021. doi:10.22337/2073-8412-2021-1-35-43.

[9] S. Bushuyev, N. Bushuyeva, D. Bushuiev, V. Bushuieva, Cognitive readiness of managing infrastructure projects driven by SMARTification. 2022 IEEE European Technology and Engineering Management Summit, E-TEMS, Conference Proceedings, 2022, pp. 196–201. doi:10.1109/E-TEMS53558.2022.9944458.

[10] S. Rudenko, T. Kovtun, Creation of the Eco-Logistic system project products configuration in the conditions of uncertainty. CEUR Workshop Proceedings, 2851, 2021, pp. 195–205.

[11] S. Chernov, L. Chernova, L. Chernova, N. Kunanets, V. Piterska, The Synergetic Effect in the Management of Active System with Distributed Control. International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2023. doi:10.1109/CSIT61576.2023.10324123.

[12] A. Shakhov, V. Piterska, V. Botsaniuk, O. Sherstiuk, Mechanisms for Goal Setting and Risk Management of Concession Projects in Seaports. International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2, 2020, pp. 185–189, 9321963. doi:10.1109/CSIT49958.2020.9321963.

[13] A. Shakhov, O. Kyryllova, O. Sagaydak, V. Piterska, O. Sherstiuk, Conceptual Risk-oriented Model of Goal Setting in the Implementation of Concession Projects in Seaports. CEUR Workshop Proceedings, 3295, 2022, pp. 149–158.

[14] S. Bushuyev, S. Onyshchenko, N. Bushuyeva, A. Bondar, Modelling projects portfolio structure dynamics of the organization development with a resistance of information entropy. International Scientific and Technical Conference on Computer Sciences and

Information Technologies, 2, 2021, pp. 293–298. doi:10.1109/CSIT52700.2021.9648713.

[15] A. Bondar, S. Bushuyev, S. Onyshchenko, H. Tanaka, Entropy paradigm of project-oriented organizations management. CEUR Workshop Proceedings, 2565, 2020, pp. 233–243.

[16] S. Rudenko, T. Kovtun, V. Smrkovska, Devising a method for managing the configuration of products within an eco-logistics system project. Eastern-European Journal of Enterprise Technologies 4 (3-118) (2022) 34–42. doi:10.15587/1729-4061.2022.261956.

[17] S. Rudenko, T. Kovtun, V. Smrkovska, Formation of the balanced development trajectory of the ecologistic system project. Eastern-European Journal of Enterprise Technologies 2(3 (122) (2023) 42–53. doi:10.15587/1729-4061.2023.277253.

[18] I. Lapkina, M. Malaksiano, Elaboration of the equipment replacement terms taking into account wear and tear and obsolescence. Eastern-European Journal of Enterprise Technologies 3(3-93) (2018) 30–39. doi:10.15587/1729-4061.2018.133690.

[19] O. Holovin, V. Piterska, Shakhov, O. Sherstiuk, Project-based Management of the Production Equipment Maintenance and Repair Information System. 3rd International Workshop IT Project Management, ITPM 2022. CEUR Workshop Proceedings, 3295, 2022, pp. 76–85.

[20] S. Bushuyev, N. Bushuyeva, V. Bushuieva, & D. Bushuiev. SMART intelligence models for managing innovation projects. CEUR Workshop Proceedings, 3171, 2022, pp. 1463–1474.

[21] S. Rudenko, T. Kovtun, T. Smokova, I. Finohenova, The genetic approach application and creation of the project genetic model. International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2022, pp. 434–437. doi:10.1109/CSIT56902.2022.10000822.

[22] Redfin. Redfin Real Estate, 2023. URL: https://www.redfin.com.

[23] Scikit-learn Developers. TfidfVectorizer - scikit-learn 0.24.2 documentation, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[24] Scikit-learn Examples. Recursive Feature Elimination with Cross-Validation, 2023. URL: https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html.

[25] M. Mao, A Comparative Study of Random Forest Regression for Predicting House Prices Using. Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023), 2024, pp. 619–626. doi:10.2991/978-94-6463-370-2_63.

[26] L. M. John, R. Shinde, S. Shaikh, D. Ashar, Predicting House Prices using Machine Learning and LightGBM. Proceedings of the 7th International Conference on Innovations and Research in Technology and Engineering (ICIRTE-2022), John College of Engineering and Management, 2022. doi:10.2139/ssrn.4108744.

[27] SHAP Documentation. SHAP Explainer Documentation. Retrieved from https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html.

[28] Scikit-learn Documentation. sklearn.inspection.PartialDependenceDisplay, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.inspection.PartialDependenceDisplay.html.