Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 207–212
DOI: 10.15439/2024F7603
ISSN 2300-5963 ACSIS, Vol. 39

Enhancing Airbnb Price Predictions with Location-Based Data: A Case Study of Istanbul

Özgün Akalın
Computer Engineering
Galatasaray University
Istanbul, Turkey
0009-0001-2067-856X

Gülfem Isiklar Alptekin
Computer Engineering
Galatasaray University
Istanbul, Turkey
0000-0003-0146-1581

***Abstract*—Airbnb, a prominent online marketplace, facilitates short- and long-term rentals by connecting customers with property owners offering entire apartments or private rooms. Accurate price prediction is crucial for both the platform and rental property owners. Previous studies have primarily focused on statistical methods and pre-processing techniques, with limited exploration of the impact of location attributes. This paper aims to enhance price prediction models for Airbnb listings by incorporating location data. Utilizing data from InsideAirbnb for Istanbul, we implemented various data pre-processing techniques and enriched the dataset with location-specific information. Our findings show that incorporating these location-based features significantly improved model performance, increasing the adjusted R^2 metric by 22.5% and reducing Mean Absolute Error (MAE) by %10. This enhancement was achieved by using location-related index values and public transportation data provided by the Istanbul Metropolitan Municipality. These advancements can help property owners optimize rental prices and assist urban planners in making informed decisions about city infrastructure development.**

***Index Terms*—Price prediction, regression, XGBoost, location features, Airbnb**

I. Introduction

AIRBNB is an online platform where hosts rent their entire apartments or rooms to guests [1]. As of the end of 2023, it serves as a global marketplace with over 5 million hosts and 7.7 million active listings. Airbnb's offerings differ from conventional hotels because each home is unique. Different needs of guests also contributed to different offerings on Airbnb, ranging from short-term to long-term rentals, each with different amenities and sizes. Since each accomodation is distinct, pricing is determined by the host. Setting the appropriate price is therefore crucial, as an excessively high price can result in missed bookings, while a low price can lead to a loss of potential income.

In this paper, we aim to enhance the performance of price prediction models for Airbnb listings by incorporating location data. Accurate price prediction is important for two stakeholders: First, it is important for the marketplace company, such as Airbnb, by enabling the provision of automated recommendation services to renters. Such services, based on statistical information, can result in a higher number of overall bookings in the system and enhance competitiveness. Second, it is crucial for rental property owners, as numerous parameters influence the price of a listing, making it challenging to determine the optimal price.

Previous research in price prediction for rental accommodations has primarily focused on utilizing various statistical methods, comparing their performances, and applying different pre-processing techniques to available datasets [2] [3] [4]. However, beyond the physical properties of the rental, the impact of location, characterized by its multiple attributes, has not been fully explored. This research contributes the following findings:

- A fine-grained analysis of the importance of proximity to different public transportation mediums, such as ferries and taxis.
- An exploration of the impact of eight distinct neighborhood metrics on rental prices, such as cultural activity and health.
- Measurements of the extent to which the location features we integrated and calculated enhance the performance of prediction models.

The structure of the paper is as follows: Section II discusses related works on price predictions for Airbnb. Section III presents our datasets (Airbnb, 34 Minutes Istanbul Index, and Istanbul Metropolitan Municipality Public Transportation) and their features. Section IV outlines our methodology. Section V presents our findings, including performance evaluation results. Finally, Section VI concludes the paper with recommendations and highlights open issues for further research.

II. Related Work

For many years, the literature has featured numerous studies aimed at predicting rental or sale prices. To identify the studies most similar to ours, we narrowed our search to the specific domain of price prediction for Airbnb. A price prediction model for the Beijing Airbnb market is proposed in [5]. The authors have used XGBoost and neural networks to predict the prices. They had the objective of finding the most significant and representative features to enhance the model's accuracy. The authors revealed that the XGBoost model performs best.

In [6], the authors explored various machine learning models to accurately predict Airbnb rental prices based on property characteristics such as type, location, customer reviews, availability, and year built. Eight regression models were

Topical area: Advanced Artificial Intelligence in Applications

tested, with four utilizing decision tree algorithms. The study found that decision tree-based models, particularly the random forests model, yielded the best results.

A paper examined Airbnb listings in New York City to develop a price prediction model using various methods, including linear regression, generalized additive models, deep neural networks, random forests, XGBoost, and bagging. The strongest performance was observed with bagging, XGBoost, and random forest models. These models provide insights into the determinants of listing prices and forecast future prices, offering valuable information for hosts, stakeholders, and the accommodation industry within the sharing economy [7].

The research of Alharbi (2023) [8] developed a sustainable price prediction model for Airbnb listings in Barcelona by incorporating property specifications, owner information, and customer reviews. The study found that Lasso and Ridge models performed best, with a R^2 score of 99%. Significant features impacting price predictions included sentiment polarity, number of bedrooms, accommodation capacity, number of beds, and recent reviews.

The customer reviews, house features, and geographical data were demonstrated to be effective predictive factors for Airbnb rentals [9]. The authors revealed that using multimodal data yields higher accuracy than single-type data. By incorporating numeric, text, and map data, the study identifies that advanced algorithms like deep neural networks and XGBoost outperform linear models such as linear regression and support vector regression.

III. Data

A. Airbnb Data

For this research, we utilized data from InsideAirbnb [10] for Istanbul dated March 31, 2024. InsideAirbnb provides comprehensive information about rental home listings on Airbnb.com. The data includes details about the physical properties of accommodations, their locations, review scores, host metrics, and price. The available information is a snapshot of listings at a time. The website also features an "Explore The Data" section, which offers summaries of the data and maps showing the locations of listings. Fig. 1 displays the distribution of listings in the central parts of Istanbul. The variables provided by InsideAirbnb and used in our price prediction models are listed in Table I. These variables include attributes such as the maximum capacity, the number of beds, bathrooms, and bedrooms, number of reviews, host-related information, number of night to stay, and amenities. This diverse range of variables allows for a robust analysis of how different factors contribute to the pricing of Airbnb listings.

B. 34 Minutes Istanbul Index Data

Released by Istanbul Metropolitan Municipality (IMM), 34 Minutes Istanbul [11] is a platform that evaluates how well each location in Istanbul meets various daily urban needs using 11 different indexes, rated on a scale from 0 to 100. Below is a summary of how each index is calculated:

TABLE I
VARIABLES USED IN INSIDEAIRBNB DATA

Variable	Description
accommodates	The maximum capacity of the listing
bedrooms	The number of bedrooms
beds	The number of bed(s)
bathrooms	The number of bathrooms in the listing
number of reviews	The number of reviews the listing has
review scores rating	Review score between 1 and 5
host identity verified	Whether host has verified identity
host is superhost	Whether host has superhost badge
host has profile pic	Whether the host has a profile picture
instant bookable	Whether the guest can automatically book the listing without the host requiring to accept their booking request
calculated host listings count	The number of listings the host has in the current scrape, in the city/region geography
minimum nights	Minimum number of night stay for the listing
amenities	Features of accommodation
host verifications	Verified information about host
room type	Type of listing (Entire home/apt, Private room, Shared room, Hotel)

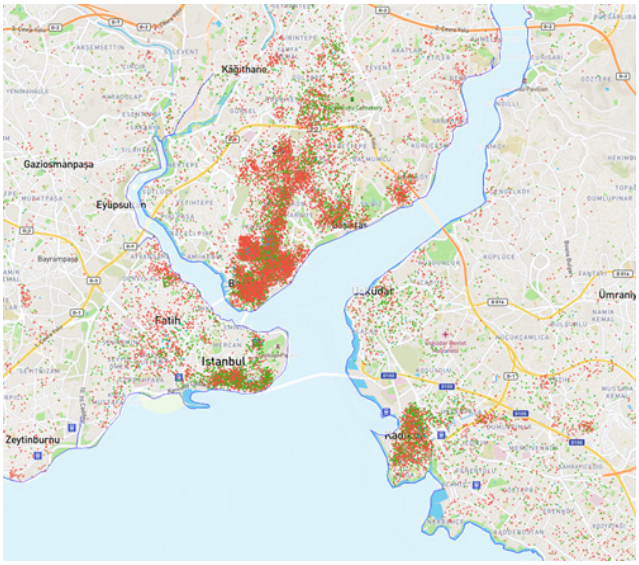

Fig. 1. InsideAirbnb data - Listings Map Visualization

- **Shelter:** Urban density, pollution, recycling, building seismicity, housing type diversity, police stations and municipal WiFi access.
- **Work:** Business centers, offices, plazas, and industrial areas.
- **Meeting Your Needs:** Stores, buffets, markets, local services, state organizations, banks, and similar facilities.
- **Cultural Activity:** Cultural centers, social facilities, the-

aters, cinemas, and other cultural venues.
- **Learning:** Nurseries, kindergartens, schools, libraries, special education, and non-formal education centers.
- **Health:** Health centers, hospitals, veterinary clinics, pharmacies, health services, and fire departments.
- **Transport:** Bike stations, bus stops, taxis, rail systems, and parking lots.
- **Spending Time:** Green spaces, monuments, squares, places of worship, gyms, cafes, and restaurants.
- **Affordability:** Housing affordability for residents.
- **Walkability:** Areas with well-maintained pedestrian paths and sidewalks that are attractive, comfortable, safe, connected to transportation, and accessible.
- **Quality of Life:** Average of diversity, affordability, an walkability indexes.

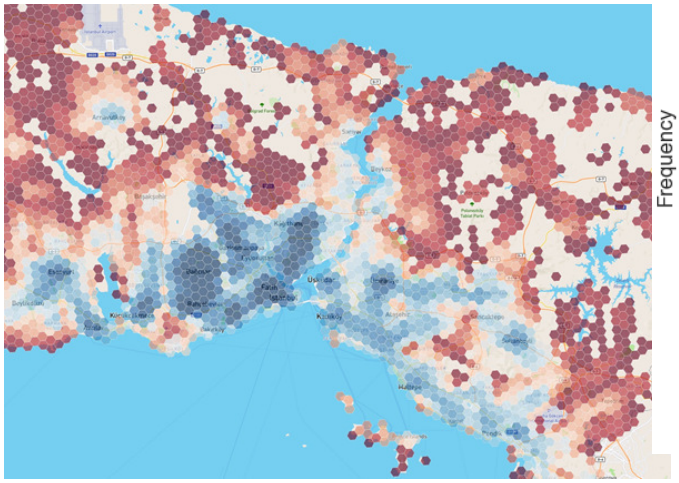

Fig. 2. Meeting Your Needs Index Across Istanbul

Fig. 2 illustrates the *meeting your needs* index across Istanbul, showcasing how different areas perform in terms of providing essential services and amenities. The 34 Minutes Istanbul data is available as both a GeoJSON file and a website that visually represents the index distribution using color codes.

C. Public Transportation Data

IMM data website [12] provides up-to-date information on public transportation. This data is categorized by transportation methods, including taxis, taxidolmus, minibuses, railways, mobility services, and ferries. It includes details such as stop names, locations, and usage statistics (where available).

For this research, we utilized the latitude and longitude data of public transport stops across each category. This spatial data was essential for enriching our dataset and improving the accuracy of our predictive models by incorporating proximity to public transportation as a key variable.

IV. Methodology

A. Data Processing and Feature Engineering

Before training the model and predicting price for Airbnb listings, certain data processing and feature engineering steps are applied on InsideAirbnb data.

The target variable, price, contained outliers and some null values. Rows with null values for price were removed, and listings with prices greater than 8000 were eliminated. This resulted in the price distribution shown in Fig. 3. After removal of the outliers, mean value of price was 1997, and standard deviation was 1380. For other columns with missing values, the empty entries were replaced with the mean value of the respective column. The amenities and host verifications were represented as lists in the InsideAirbnb data; therefore, the lengths of these lists were calculated and stored in two new columns, *numberOfAmenities* and *numberOfHostVerifications*.

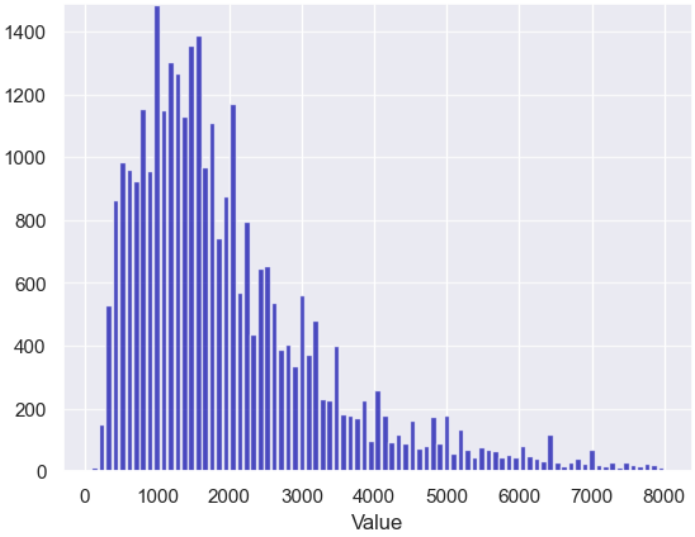

Fig. 3. Distribution of Target Variable Price After Pre-Processing

Columns containing 't' (true) and 'f' (false) values, such as *instant bookable*, *host has profile pic*, *host is superhost*, and *host identity verified*, were transformed into binary values (1 and 0). The *room type* column, which had four distinct values, was one-hot encoded, resulted in four new columns: *Entire home/apt*, *Hotel room*, *Private room*, and *Shared room*.

For each listing, latitude and longitude information were used to determine values for 11 different 34 Minutes Istanbul indexes at each listing's location, as well as the number of stops for various public transportation categories within 0.5, 1, 3, and 5 kilometers.

Correlation analysis revealed that mobility stops (parking areas for rentable scooters and bicycles) and taksidolmus stops did not significantly affect the price and were excluded from the model training. For the remaining categories (taxi, minibus, railway and ferry) the information most correlated with price is presented in Table II. Additionally, the *Spending Time*, *Work*, and *Transport* indexes were removed due to their low correlation with the target variable.

After removing outliers, the final dataset for training the models comprised 31 columns and 32,555 rows. Fig. 4 shows the variables with a correlation greater than 0.05 with the target variable, price.

TABLE II
DISTANCE USED FOR CALCULATING NUMBER OF STOPS

Transportation Method	Top Correlated Distance
Ferry	1 km
Railway	0.5 km
Taxi	0.5 km
Minibus	1 km

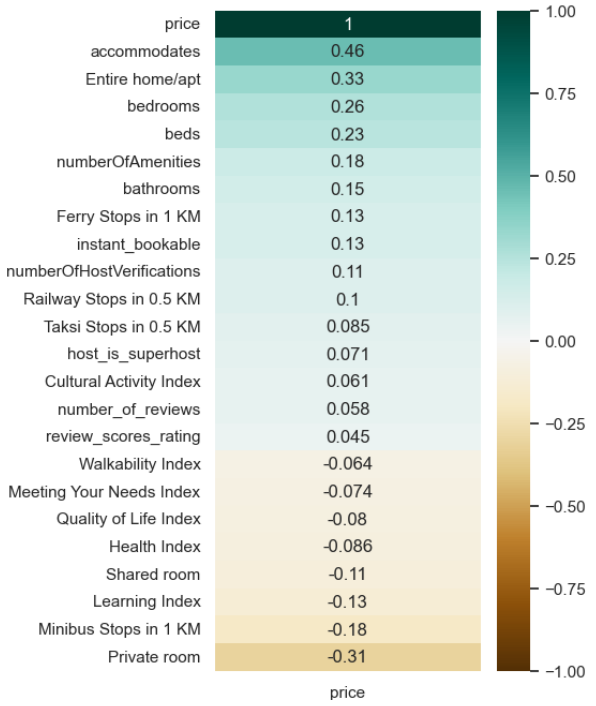

Fig. 4. Correlation of Features with Target Variable

B. Regression Models

We employed linear and XGBoost [13] [14] regression models to predict Airbnb listing prices. The dataset was randomly split into two subsets: 80% for training and 20% for testing. Same listings were used for training and testing groups across all experiments. Each model was trained on the training data and subsequently used to predict the prices of listings in the test data. We conducted our experiments using base InsideAirbnb data, and added location based features, in different combinations.

C. Evaluation Metrics

In order to compare the performance of different regression methods and feature sets, we used MAE and adjusted R^2 as evaluation metrics.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$$

MAE, as described by Sammut et al. (2010) [15], is the average of absolute errors $|y_i - \widehat{y}_i|$, where y_i is the actual value and $\widehat{y}_i$ is the predicted value.

Adjusted R^2, an extension of the coefficient of determination R^2 proposed by Wright (1921) [16], is calculated as:

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{N-1}{N-k-1}$$

where N is the sample size and k is the number of predictor variables. R^2 measures how well the model fits the data.

$$R^2 = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$

Adjusted R^2 [17] is preferred over R^2 in this context because it accounts for the number of predictor variables in the model. This adjustment is particularly important when comparing models trained with different feature sets, such as those that include location-based variables and those that use only the InsideAirbnb variables. Adjusted R^2 penalizes the inclusion of additional predictor variables, addressing the issue highlighted by Miles (2005) [17], where R^2 always increases with each added variable, potentially misleading the evaluation of model performance.

V. RESULTS

Experiment results both with and without location data show that XGBoostRegressor outperforms Linear Regression significantly. Using base Airbnb data, the adjusted R^2 was 0.258 for the linear regression model, and 0.403 for the XGBoost regression model, with mean absolute error 858 and 747, respectively.

The prediction performances of both models improved significantly after adding location-based variables. Experiments were conducted by training and testing models with three different combinations of feature sets. Firstly, we added data on the proximity to four different public transportation mediums. Secondly, we utilized neighborhood information categorized into eight distinct values. Lastly, we considered a combination of both feature sets.

The results demonstrated that incorporating public transportation data alone increased the adjusted R^2 metric from 0.403 to 0.448. In comparison, using neighborhood index data alone increased it to 0.468.

Using both feature sets increased the adjusted R^2 metric from 0.258 to 0.306 for the linear regression model and from 0.403 to 0.491 for XGBoost regression model. Additionally, for the XGBoost regressor, the MAE decreased from 747 to 672 when all the predictor variables were included.

TABLE III
MODEL PERFORMANCE METRICS

	Linear Regr.		XGBoost	
Features	**MAE**	**Adj.R^2**	**MAE**	**Adj.R^2**
Airbnb	858	0.258	747	0.403
Airbnb + Public Transport(PT)	836	0.286	712	0.448
Airbnb + 8 Indexes	830	0.295	693	0.468
Airbnb + PT + 8 Indexes	820	0.306	672	0.491

In order to measure the impact of new features individually, we calculated their permutation importances. For public

transportation features, we used the model trained with Airbnb and IMM Public Transport features. For neighborhood index features, we used the model trained with Airbnb and the eight index features. The adjusted R^2 metric was used as the scoring method. Fig. 5 illustrates the impact of each index feature in the model. Among the eight indexes, the cultural activity index and learning index had the most significant impact on the model performance. They are followed by the meeting your needs and walkability indexes. The other four features had lower significance, with the health index being the least impactful.

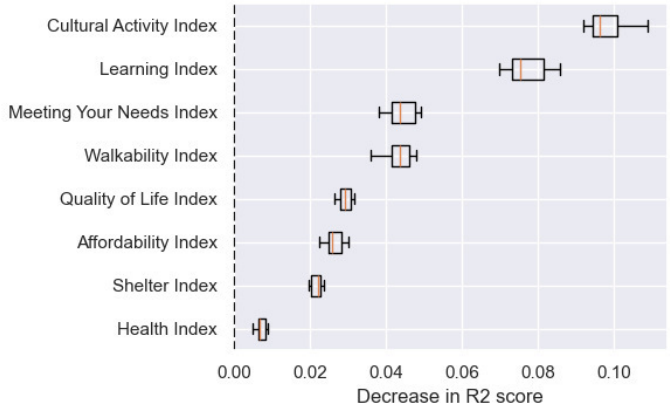

Fig. 5. Permutation Importance of Location Indexes

As for the effect of a rental's proximity to nearby public transportation stops in different categories, Fig. 6 shows that an accommodation's being close to minibus stops within 1 kilometer is the leading contributing factor, followed by being next to ferry and taxi stops. Proximity to railway stops contributed the least to the performance of the model.

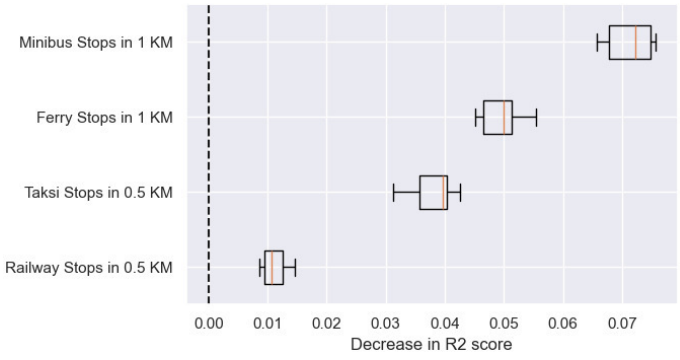

Fig. 6. Permutation Importance of Public Transportation Proximity

VI. Conclusion

Our analysis has demonstrated that the price of an Airbnb listing is influenced by a multitude of parameters, with the physical properties of the accommodation being the most significant. This is followed by location-related factors, as well as host and listing-specific information.

Among the regression models tested, XGBoost exhibited the best performance, corroborating the findings of prior research [9] [5]. Peng et al. (2020) [9] found an average R^2 score of 0.477 with XGBoost using data from 10 different cities, whereas Yang (2021) [5] found an R^2 of 0.655 for Beijing alone. Our results for Istanbul using only Airbnb data resulted in a value of 0.403. This initially lower fit could be due to fewer data points or the absence of certain features in the InsideAirbnb Istanbul data, such as cleaning and security fees.

By enriching the dataset with location-based data, such as the cultural activity index, quality of life index, and the number of public transportation stops within specified distances, we observed a substantial improvement in model performance. Specifically, the adjusted R^2 metric for XGBoost increased by approximately 22%, from 0.403 to 0.491. Compared to other studies that used location-based features [18] [19] [20] and showed improvements relative to the base model, our study revealed one of the highest increases in both absolute and relative performance. For example, Schwarzová (2020) [18] increased the R^2 value from 0.623 to 0.628 by adding crime-related neighborhood data, Chica-Olmo et.al (2020) [19] increased the R^2 value from 0.363 to 0.446, and Luo and Kawabata (2018) [20] from 0.441 to 0.498.

By calculating the individual importance of each location metric we added to the dataset, we showed that cultural activity, learning, meeting your needs, and walkability characteristics of a neighborhood play a role in the determination of the pricing. Although cultural activity is no surprise due to its inherent relationship with touristic activities, the learning index that is calculated by taking schools, libraries, etc. into consideration was not observed in previous studies. Further research is needed on whether similar metrics play a contributing role to prices in other cities around the world. We also showed that different proximity to different types of transportation stations affect the price differently.

A. Threats To Validity

The data available at InsideAirbnb includes information about listings at a particular time. As a result, price changes over a span of time cannot be investigated and taken into account. Also, it is unknown whether prices of properties provided in the dataset are affected by seasonality or other events that change supply and demand. Daily price information for each listing over the course of months instead of a single day would allow time-series regression.

B. Future Work

Future work will build upon our current findings to explore the impact of similar location-based data enrichment on the prediction of Airbnb rental prices in other cities globally. Although the data provided by IMM is specific to Istanbul, we aim to generalize the underlying principles of key indexes to develop a broadly applicable solution.

Additionally, we plan to incorporate the usage frequency of public transportation stops by assigning weights based on the number of users. This approach will allow us to differentiate

between the effects of stops with varying levels of usage, potentially enhancing the precision of our predictive models.

ACKNOWLEDGMENT

This research has been financially supported by Galatasaray University Research Fund, with project ID: FBA-2024-1258.

REFERENCES

[1] Airbnb: https://news.airbnb.com/about-us/
[2] Ghosh, Indranil, Rabin K. Jana, and Mohammad Zoynul Abedin. "An ensemble machine learning framework for Airbnb rental price modeling without using amenity-driven features." International Journal of Contemporary Hospitality Management 35.10 (2023): 3592-3611.
[3] Kirkos, Efstathios. "Airbnb listings' performance: Determinants and predictive models." European Journal of Tourism Research 30 (2022): 3012-3012.
[4] Wang, Haoqian. "Predicting Airbnb listing price with different models." Highlights in Science, Engineering and Technology 47 (2023): 79-86.
[5] Yang, Siqi. "Learning-based Airbnb price prediction model." 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). IEEE, 2021.
[6] Lektorov, A., Abdelfattah, E., and Joshi, S. "Airbnb Rental Price Prediction Using Machine Learning Models," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0339-0344.
[7] Zhu, A., Li, R., and Xie, Z. "Machine Learning Prediction of New York Airbnb Prices," 2020 3rd International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 2020, pp. 1-5.
[8] Alharbi, Z.H. "A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis." Sustainability 2023, 15, 13159.
[9] Peng, Ningxin, Kangcheng Li, and Yiyuan Qin. "Leveraging multimodality data to Airbnb price prediction." 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME). IEEE, 2020.
[10] InsideAirbnb: https://insideairbnb.com/get-the-data
[11] 34 Dakika Istanbul: https://34dakika.istanbul/map
[12] Istanbul Metropolitan Municipality https://data.ibb.gov.tr/en/dataset
[13] Lewandowska, Alexandra. "XGBoost meets TabNet in Predicting the Costs of Forwarding Contracts," 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS) (2022): 417-420.
[14] Podlodowski, Ł. and Kozłowski, M. "Predicting the Costs of Forwarding Contracts Using XGBoost and a Deep Neural Network," 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS) (2022): 425-429.
[15] Sammut, Claude, and Geoffrey I. Webb. "Mean absolute error." Encyclopedia of Machine Learning 652 (2010).
[16] Wright, Sewall. "Correlation and causation." Journal of Agricultural Research 20.7 (1921): 557.
[17] Miles, Jeremy. "R-squared, adjusted R-squared." Encyclopedia of Statistics in Behavioral Science (2005).
[18] Schwarzová, Lucie. Predicting Airbnb Prices with Neighborhood Characteristics: Machine Learning Approach. Diss. Tilburg University, 2020.
[19] Chica-Olmo, Jorge, Juan Gabriel González-Morales, and José Luis Zafra-Gómez." Effects of location on Airbnb apartment pricing in Málaga." Tourism Management 77 (2020): 103981.
[20] Luo, Yanjie, and Mizuki Kawabata. "Airbnb pricing and neighborhood characteristics in San Francisco.", Available at: https://tinyurl.com/w53w277s, 2018.