Real estate price prognostication through machine learning models

Doddamani John Vestly
Department of Computer Science and Engineering, Vardhaman College of Engineering
Hyderabad, India
doddamanijohn123@gmail.com

Neelima Thangallapelly
Department of Computer Science and Engineering, Vardhaman College of Engineering
Hyderabad, India
neelimathangallapelly@gmail.com

Gandra Sai Manish
Department of Computer Science and Engineering, Vardhaman College of Engineering
Hyderabad, India
manishgandra7485@gmail.com

Shanmugasundaram Hariharan
Department of Computer Science and Engineering, Vardhaman College of Engineering
Hyderabad, India
mailtos.hariharan@gmail.com

Vinay Kukreja
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
vinay.kukreja@chitkara.edu.in

H. Venkateswara Reddy
Department of Computer Science and Engineering, Vardhaman College of Engineering
Hyderabad, India
h.venkateswarareddy@vardhaman.org

***Abstract*— In the dynamic realm of real estate, house prices became the crucial part to encounter the set of hurdles to make lives better. Through meticulous feature engineering, which includes advanced data cleansing and feature manipulation, coupled with robust machine learning techniques, our model offers a nuanced understanding of valuation dynamics. The optimization process involves hyper parameter tuning and cross-validation, employing cutting-edge methodologies to extract latent patterns and yield meaningful insights from the underlying data. Leveraging algorithms of supervised learning algorithms like linear regression and K-fold, chosen for their ability to discern intricate patterns within diverse datasets, our research pioneers a transformative approach to real estate valuation. Evaluation metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and R-squared were used to ensure a robust and accurate predictive framework which were promising**

Keywords*— *Real Estate, Data Science, Machine Learning, Predictive Modeling, Feature Engineering, Valuation, Algorithms.

I. Introduction

Determining home values in the ever-changing real estate market is a complex task with far-reaching consequences for people, neighborhoods, and larger economic systems. This task's intrinsic complexity highlights the necessity for a methodical, data-driven approach that goes beyond conventional valuation techniques to provide a comprehensive knowledge of the complex factors impacting real estate pricing [1]. Fundamental to this study is the understanding that accurate housing price prediction necessitates a deviation from traditional approaches [2].

The core issue statement centers on the necessity of accurately predicting home prices while taking into account a variety of factors and changing market conditions. In order to meet this difficulty, our model uses a laborious process called feature engineering, which entails sophisticated data cleansing and manipulation techniques. Supported by strong machine learning techniques, the model is not only built to identify complex patterns in the data but is also optimized via cross-validation and hyper-parameter tweaking [3]. The fundamental conundrum is that property values are dynamic, impacted by a several factors depending upon market conditions. In order to tackle this, our model carefully designs features by utilizing cutting-edge methods for data cleaning and processing [4].

Along with revealing hidden patterns in the data, this strategy makes use of strong machine learning techniques to maximize prediction accuracy through cross-validation and hyper-parameter optimization [5]. The significance of this research lies in its commitment to address these challenges through a synergy of data science and web development. Real estate is a sector that thrives on information and informed decisions. In this regard, data science, with its powerful predictive capabilities, offers a new paradigm for understanding and demystifying property valuations. Simultaneously, web development empowers users to interact with predictive models in real time, thus creating an accessible, transparent, and user-centric platform for navigating the intricacies of the real estate market [6].

Earlier existing project that exists not merely an academic exercise but a practical solution to a real-world problem. It seeks to empower property seekers with the tools and information necessary to make informed decisions [7]. By doing so, it aims to level the playing field, making property valuations more accessible, affordable, and transparent for a broader spectrum of individuals. This paper outlines the approach we have taken, which combines advanced data analysis, machine learning, and user-centric web

979-8-3503-4846-0/24/$31.00 ©2024 IEEE

development to offer a comprehensive, customizable, and informative solution to property valuation [8].

In the subsequent sections, we will delve into the comprehensive methodology employed in our research, presenting the data collection and preprocessing, feature engineering, machine learning techniques, and web development processes in detail. We will also present the results and insights gained through our approach, showcasing the accuracy of property price predictions and the positive user experiences with our web application. In conclusion, we will reflect on the contributions of our research, suggest potential avenues for future work, and underline the transformative potential of technology in the realm of real estate.

II. Related Work

Data science and machine learning applications in real estate has gained traction as a means to enhance property valuation accuracy. Previous research has demonstrated the effectiveness of predictive modeling in estimating property values [1]. Additionally, web development has played a crucial role in creating user-friendly interfaces for property seekers. The realm of real estate valuation, data science, and web development has witnessed a steady evolution, driven by a combination of technological advancements and the growing demand for accessible, transparent, and user-centric solutions. This literature survey explores key insights from these domains, providing a foundation for our research and highlighting the gaps that our project aims to fill [2].

The valuation of real estate properties has been a subject of research and interest for several decades [3]. The property evaluation process involves appraisers who consider various factors which includes property, location, condition, size, market trends, comparable property sales etc in order to determine the fair market value. However, this approach can be subjective, time-consuming, and may not always provide the level of transparency desired by property seekers [4].

Several studies have explored the determinants of property prices. For example, studies have found that proximity to essential amenities, such as schools, transportation hubs, and shopping centers, can significantly impact property values [5]. Research in this field has also emphasized the role of property attributes, such as square footage, number of bedrooms, and property age, as key factors influencing valuations. The application of data science and machine learning in real estate has gained prominence in recent years. These techniques have the potential to offer more accurate and data-driven property valuations [6]. Researchers have employed regression models, decision trees, and ensemble methods to predict property prices with a varying degree as compared to traditional appraisal methods [7].

Machine learning models can process large datasets, extract complex patterns, and make predictions based on a wide range of features [8]. Moreover, they can adapt to changing market conditions, providing a dynamic and up-to-date approach to property valuation. The ability to harness the power of data to understand the intricate relationships between variables and property prices is a significant leap forward in the real estate sector [9].

The growing integration of technology into the real estate sector is evident in the proliferation of online property listings, virtual tours, and real-time market data. However, these technologies often focus on presenting existing property listings rather than assisting users in assessing the value of properties or exploring the relationship between property attributes and prices [10]. Our research project diverges from the norm by emphasizing prediction and user interaction. It creates a dynamic and responsive web application that harnesses the power of data science to predict property prices based on specific criteria [11].

There are ethical questions when data science is used to value real estate. Algorithm bias, fairness, and transparency issues have been investigated. Scholars stress the significance of moral behavior while creating models to guarantee fair and reasonable results in real estate decision-making [12]. Both traditional and data-driven models are challenged by problems such data heterogeneity, inadequate information, and market volatility. To tackle these obstacles, creative solutions and a thorough comprehension of the complexities involved are needed [13].

III. Methodology

A. *Data Collection:*

The well-known dataset repository known as Kaggle is where the dataset used in this study was obtained. The dataset includes important columns that are essential for real estate value analysis, offering a full range of elements to investigate and precisely forecast property prices. The following columns are included in it: "area_type," "availability," "location," "size," "society," "total_sqft," "bath," "balcony," and "price." All of these columns combined offer a full range of features required for real estate price analysis and forecasting:

.

B. *Data pre-processing:*

Thorough pre-processing was performed on the dataset, which included imputation to fill up the gaps for the variables "availability," "balcony," and "society." One-hot encoding and label encoding were used to encode categorical variables such as "area_type" and "location." To ensure consistency for further research, numerical features most notably 'total_sqft' were also normalized using Standard and Min-Max scaling.

C. *Basic algorithm and background:*

i. Linear regression:

A fundamental supervised learning approach for predicting a continuous output variable based on one or more input data is called linear regression. The inputs and outputs of the model are assumed to have a linear relationship, which is represented by a straight line in a multidimensional space. The objective is to identify the best-fit line, usually shown by metrics like Mean Squared Error (MSE), that minimizes the deviation between the expected and actual values. Widely used in regression assignments, linear regression is an interpretable, computationally efficient approach that offers insights into the contribution of each feature to the anticipated outcome.

ii. Random forest:

Using the power of several decision trees, Random Forest is an ensemble learning approach that improves prediction accuracy. A random subset of characteristics and data samples are used to build each tree in the forest, and the combined forecasts of all the individual trees yield the final prediction. By using an ensemble technique, overfitting is reduced and intricate correlations in the data are captured. The robustness, adaptability, and capacity to handle both category and numerical data are well-known attributes of Random Forest. It also offers insightful information about the significance of features, which helps analyze the behavior of the model.

iii. Support vector machine (SVM):

A flexible supervised learning technique for classification and regression applications is called Support Vector Machine (SVM). SVM seeks to identify the hyperplane in the regression context that most accurately depicts the connection between the input characteristics and the target variable. It functions by locating support vectors, or data points that affect the hyperplane's position. With the introduction of kernel functions, SVM can handle both linear and non-linear interactions and is efficient in high-dimensional areas. Because of their durability and adaptability, support vector machines (SVMs) are widely used in real estate valuation for both prediction tasks and diverse datasets.

D. Evaluation metrics:

i. R-squared error:

A statistical metric called R-squared, or coefficient of determination, is used to assess how well a regression model fits data. It sheds light on the percentage of the dependent variable's variance that the model's independent variables account upon. R-squared measures calculates the proportion of variation in the targeted variable that the models the system in proper order.

$$R^2 = 1 - SSR/SST \qquad (1)$$

ii. *MAPE (Mean Absolute Percentage Error):*

A popular statistic for assessing the precision of regression or forecasting models is the Mean Absolute Percentage Error (MAPE), especially in situations when the size of mistakes is significant. The significant difference between expected and actual values is expressed using MAPE, which gives a clear picture of the model's relative performance.

$$MAPE = 100/n \ \Sigma_{(i=1)}^{n} \ |(Y_i-\hat{Y})/Y_i| \qquad (2)$$

Where: n = Number of observations
$\hat{Y}$ = Predicted Values
Y_i= Actual Values

ii. *MAE (Mean Absolute Error):*

A simple statistic for assessing how well a regression model predicts the future is the Mean Absolute Error (MAE). It provides a clear picture of the model's magnitude performance by measuring the average of absolute difference between the predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}\right| \qquad (3)$$

$wher$: n = Number of observations
$\hat{Y}$ = Predicted values
Y_i= Actual values

E. Architecture diagram:

Fig.1 presents the flow chart depicts a comprehensive machine learning workflow, an iterative cycle that starts with loading and diligently preparing data to fuel the model's learning. The model then embarks on a journey of discovery, iteratively analyzing the data to unveil hidden patterns and relationships. This acquired knowledge is then put to the test on unseen data, evaluating the model's ability to perform in the real world. If deemed satisfactory, the trained model is unleashed into production, ready to tackle real-world challenges.

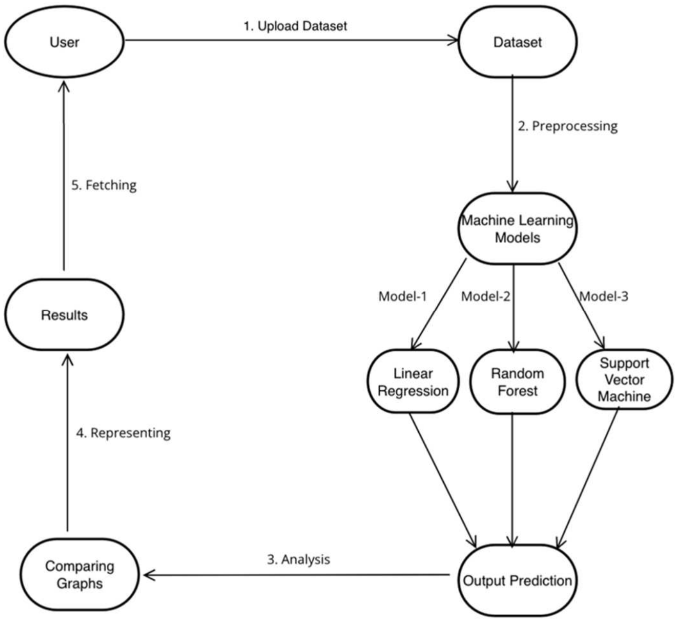

Fig-1 Architectural Flow of the Process.

However, the journey doesn't end there. The model's performance is continuously monitored and meticulously refined, ensuring its effectiveness remains optimal as new data emerges and circumstances evolve. This cyclical process of learning, evaluation, and refinement lies at the heart of successful machine learning endeavors.

TABLE I. EVALUATION METRICS OF DIFFERENT MODELS

Model	MAE	MAPE	R-Squared
Linear Regression	17.89	0.22	0.78
Random Forest	18.19	0.20	0.77
Support Vector Machine	21.74	0.23	0.69

Table 1 provides an overview of the evaluation measures for three different regression models: Support Vector Machine, Random Forest, and Linear Regression. Key performance measures such as Mean Absolute Error (MAPE), Mean Absolute Percentage Error (MAPE), and R-Squared were used to evaluate the models. The Linear Regression model performed well when it came to predicting real estate prices (presented in Fig.2). It had a low MAE (17.89), a minimal MAPE (0.22), and a high R-Squared value (0.78), all of which pointed to a good explanatory power. With a little higher MAE (18.19) and a lower MAPE (0.20), the Random Forest model showed competitive accuracy, yielding an R-Squared value of 0.77 (presented in Fig.3). In contrast, the Support Vector Machine demonstrated a greater MAE (21.74) and MAPE (0.23) with an R-Squared value of 0.69, albeit still yielding reasonable results. These results highlight the necessity of striking a compromise between interpretability and prediction accuracy, which provides insightful guidance for model selection in real estate applications.

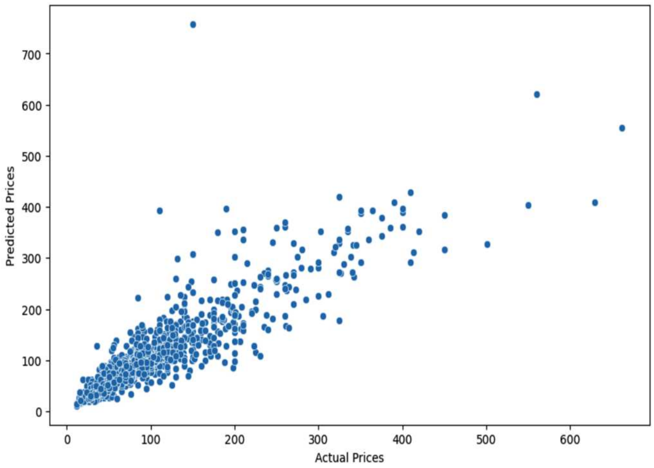

Fig.2 Linear regression performance.

IV.Result And Discussion

The work ccompleted illustrates a number of outcomes. The information about the evaluation measures and the visual element is also included in this results section.

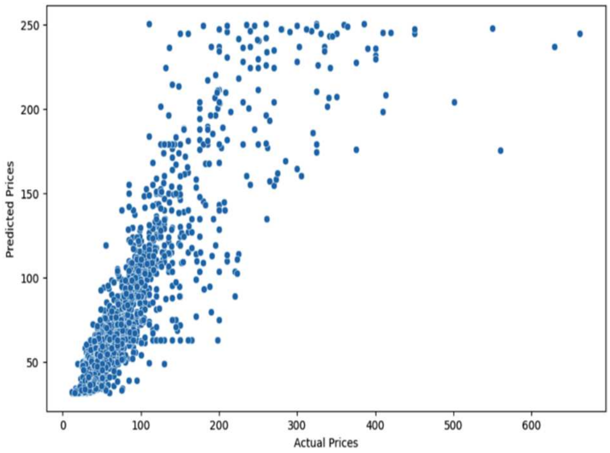

Fig-3 Random Forest performance.

The performance of different algorithm is presented in Fig.4.
.

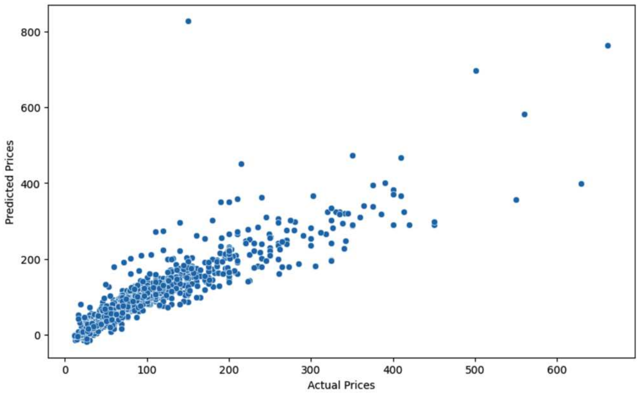

Fig. 4. Performance Metrics for Different Algorithms

The scatter plot that is being displayed shows the relationship between the actual product prices (X-axis) and the prices that a random forest model predicts (Y-axis). The data points are closely clustered around the diagonal line, demonstrating a high positive correlation and the model's efficacy in accurately predicting prices, particularly in the mid-range. Because it uses many trees trained on different subsets of data, the ensemble aspect of the random forest improves predictive strength and allows it to manage outliers and capture non-linear correlations with ease. Although there are occasional significant deviations, the general trend highlights the model's strong performance. This visual analysis offers insightful information about the advantages and disadvantages, directing efforts to optimize and modify for improved forecast accuracy in the future.

The association between actual product prices (X-axis) and prices predicted by a random forest model (Y-axis) is shown in the scatter plot that is being presented. The data points closely cluster around the diagonal line that represents perfect prediction, indicating a significant positive correlation. This shows that the model is good at correctly predicting prices, especially in the middle range. Although there are few outliers at the extremes, the general trend highlights the model's potential. The ensemble aspect of the random forest accounts for this prediction strength. It uses several decision trees, each trained on a different sample of data, in contrast to single decision trees, and when averaged together, produces a forecast that is more reliable and accurate. This improves the model's performance even further by allowing it to manage outliers and capture non-linear relationships more effectively than simpler models.

V.CONCLUSION

The study presented here investigated on the use of Support Vector Machine, Random Forest, and Linear Regression models for real estate price prediction. The base was laid via meticulous preparation, which included data cleaning and outlier reduction. Evaluation metrics demonstrated the balanced performance of Linear Regression, while visual analysis demonstrated the robust predictive strength of the Random Forest model. The results highlight the fine balance that must be struck between interpretability and accuracy. Linear regression was found to be a reasonably balanced option, while Random Forest demonstrated exceptionally high accuracy. Although reasonable, Support Vector Machine displayed more mistakes. This study provides information to help with decision-making in the ever-changing housing market. Subsequent improvements can concentrate on exploring features through feature engineering, improving the model, and using cutting-edge methods to improve prediction accuracy.

REFERENCES

[1] Z. F. Abut, H. Ş. Arlı, M. F. Akay and Y. Adıgüzel, "A New Hybrid Approach for Real Estate Price Prediction Using Outlier Detection, Feature Selection, and Clustering Techniques," 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 1-6, doi: 10.1109/UBMK59864.2023.10286673.

[2] M. Arivukarasi, A. Manju, R. Kaladevi, S. Hariharan, M. Mahasree and A. B. Prasad, "Efficient Phishing Detection and Prevention Using Support Vector Machine (SVM) Algorithm," *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, 2023, pp. 545-548, doi: 10.1109/CSNT57126.2023.10134735.

[3] S. Prongnuch, S. Sitjongsataporn, K. Intawichai and J. R. Kunkar, "Outcome-based Learning in Online STEM Activities for Robot and Real Estate Management Camp," 2022 7th International STEM Education Conference (iSTEM-Ed), Sukhothai, Thailand, 2022, pp. 1-4, doi: 10.1109/iSTEM-Ed55321.2022.9920828.

[4] A. Wandhe, L. Sehgal, H. Sumra, A. Choudhary and M. Dhone, "Real Estate Prediction System Using ML," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-4, doi: 10.1109/ICETET-SIP58143.2023.10151561.

[5] A. Peter, A. A. Kumar, A. Rajeev, B. Baiju and V. S. Chooralil, "Real Estate Management System using Blockchain," 2023 International Conference on Innovations in Engineering and Technology (ICIET), Muvattupuzha, India, 2023, pp. 1-4, doi: 10.1109/ICIET57285.2023.10220623.

[6] M. S. Bennet Praba, U. Rajeev, A. Rathore and A. Kolangarath, "Real Time Automation on Real Estate using API," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichy, India, 2022, pp. 1-5, doi: 10.1109/ICEEICT53079.2022.9768428.

[7] Y. Zheng, B. Yang, R. Zhang, Z. Bai and Y. Sun, "Mass Appraisal of Real Estate Prices Using Improved BP Neural Network with Policy Evaluation," 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 2022, pp. 1036-1041, doi: 10.1109/TOCS56154.2022.10015915.

[8] Y. Sun and G. Peng, "Developing Area Real Estate Valuation Based on Linear Regression and KNN Algorithm," 2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 2022, pp. 38-42, doi: 10.1109/ICDSBA57203.2022.00014.

[9] S. Thokala, R. Jakkani, M. Alli and H. Shanmugasundaram, "Accident Prevention System using Machine Learning," *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128202.

[10] H. A. V. P. U. Hapuarachchi, M. D. Manoratne, K. G. B. K. Gamlath, V. S. G. G, D. Sriyaratna and N. H. P. R. Supunya, "Realty Scout Smart System for Real Estate Analysis & Forecasting with Interactive User Interface," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9825335.

[11] W. Coleman, B. Johann, N. Pasternak, J. Vellayan, N. Foutz and H. Shakeri, "Using Machine Learning to Evaluate Real Estate Prices Using Location Big Data," 2022 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2022, pp. 168-172, \ doi: 10.1109/SIEDS55548.2022.9799393.

[12] Y. Zhao, X. Shen, X. Xu and Y. Xu, "Application of BP neural network in real estate batch assessment," 2022 41st Chinese Control Conference (CCC), Hefei, China, 2022, pp. 7018-7023, doi: 10.23919/CCC55666.2022.9901571.

[13] Y. Zhao, G. Chetty and D. Tran, "Deep Learning for Real Estate Trading," 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2022, pp. 1-7, doi: 10.1109/CSDE56538.2022.10089222.

[14] S. Wang, J. Zhu, Y. Yin, D. Wang, T. C. Edwin Cheng and Y. Wang, "Interpretable Multi-Modal Stacking-Based Ensemble Learning Method for Real Estate Appraisal," in IEEE Transactions on Multimedia, vol. 25, pp. 315-328, 2023, doi: 10.1109/TMM.2021.3126153.

[15] T. Zhu, "The Impact of Non-immersive Virtual Reality Technologies on Consumers' Behaviors in real estate: A Website's Perspective," 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Singapore, Singapore, 2022, pp. 13-20, doi: 10.1109/ISMAR-Adjunct57072.2022.00013.

[16] L. Kong et al., "When permissioned blockchain meets IoT oracles: An on-chain quality assurance system for off-shore modular construction manufacture," 2022 IEEE 1st Global Emerging Technology Blockchain Forum: Blockchain & Beyond (iGETblockchain), Irvine, CA, USA, 2022, pp. 1-6, doi: 10.1109/iGETblockchain56591.2022.10087164.

[17] R. Henker, D. Atzberger, W. Scheibel and J. Döllner, "Real Estate Tokenization in Germany: Market Analysis and Concept of a Regulatory and Technical Solution," 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), Dubai, United Arab Emirates, 2023, pp. 1-5, doi: 10.1109/ICBC56567.2023.10174954.

[18] T. Balasooriya et al., "Location Intelligence Based Smart E-Commerce Platform for Residential Real-Estate Industry," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 867-873, doi: 10.1109/ICOSEC54921.2022.9952023.

[19] M. Selim, M. R. Rabbani and A. Bashar, "Qard Hasan Based Cooperative Model for Home Financing and Its Effects in Home Ownership and Real Estate Development," 2022 International Conference on Sustainable Islamic Business and Finance (SIBF), Sakhir, Bahrain, 2022, pp. 48-52, doi: 10.1109/SIBF56821.2022.994002.

[20] G. Wang, J. S. Suroso, D. Sanusi, J. A. Tanuwijaya and T. F. I. Theodora, "Applying Internet of Things Framework in Real Estate Business with Enterprise Architecture Approach," 2022 International Conference on Information Management and Technology (ICIMTech), Semarang, Indonesia, 2022, pp. 219-224, doi: 10.1109/ICIMTech55957.2022.9915151.