BCT Journal

Journal of Business, Communication & Technology

Machine Learning Impact on Modern Business Intelligence

Raziyeh Moghaddas[1]*, Farinaz Tanhaei[2], Maryam Almoqbali[1], Solmaz Safari[2]

[1]Gulf College, Oman, [2]Swansea University, UK

Abstract The incorporation of machine learning (ML) approaches into business intelligence (BI) results in a great impact on fields that require predictive analysis, such as the real estate market. An accurate prediction of housing prices can provide benefits to stakeholders such as developers, investors, and policy planners. This study aims to explore the application of ML techniques to property valuation by creating a dataset based on real-world house price data collected from various areas in Muscat, Oman. Several ML models, including Linear Regression, Ridge Regression, Gradient Boosting, Random Forest, and Support Vector Regression, were applied and examined on the created dataset to estimate the house prices. Besides, hyperparameter tuning is used for each model in order to improve their predictive accuracy. Finally, we assessed the performance of each model using standard evaluation metrics, i.e., Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared (R^2) score. The findings of this research work provide a comparative analysis of model efficiency that highlights both the capabilities and limitations of each model. This study demonstrates the practical power of ML techniques in real-state analytics and its wider applicability in improving BI systems subsequently.

Keywords: *Artificial intelligence, Machine learning, House price predictor, Business intelligence, Linear regression*

1. Introduction

Artificial intelligence (AI) was introduced as a substitution for human intelligence by developing computer systems that are able to think like a human, act like a human, and learn and get experiences like a human (Russell & Norvig, 2020). Machine learning (ML) is a subset of AI that refers to the idea that computer systems can learn from and adapt to new data without the need for human intervention. Over the past years, AI and its subsets, such as ML, have demonstrated great ability in improving various operational processes, including business and marketing trends, because of its varied capabilities, like intelligent data analysis and predictions. One of the impactful applications of ML is price prediction, where it offers businesses improved accuracy in anticipating market trends and making data-informed decisions.

***Corresponding Author:**
Raziyeh Moghaddas
raziyeh@gulfcollege.edu.om

Received: March 2025
Revised: May 2025
Accepted: May 2025
Published: June 2025

© 2025 Moghaddas, Tanhaei, Almoqbali, and Safari.
This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Generally, the use of ML in business intelligence (BI) has a significant impact on the way that organizations interact with data. Modern BI is a collection of tools and approaches ranging from data collection and data mining to advanced analytics and ML that help organizations optimize their decision-making processes (Hamzehi & Hosseini, 2022; Kasemsap, 2016). Among the businesses, real estate has shown the least transparency. House prices change every day and sometimes are hyped rather than being based on actual valuation. They have caused unique challenges in complex markets such as Muscat, Oman, due to low transparency, varied neighborhood profiles, and economic volatility (Varma et al., 2018). While coping with such challenges using the traditional approaches could not provide much reliable insights, ML models are able to solve these issues by managing the interaction between variables in high dimensional data. For instance, ML ensemble models such as Random Forests have consistently shown greater predictive power compared to traditional statistical approaches (Huang et al., 2025).

However, due to the limited access to clean transactional data, weak demand, and oversupply, modeling the real state markets such as Muscat stayed challenging. Furthermore, some other economic factors, such as changes in the price of oil and even the COVID-19 pandemic's aftermath, increase the issues. Hence, more equipped analytical techniques are required to handle such instability than traditional tools. The use of BI in real estate is still at its base level, though Oman has made steps toward digital innovation through the development of smart cities and e-government programs. Recent research emphasizes the AI-based model's potential to support the property valuation's accuracy in areas with social variety and economic circumstances (e.g., Alsaidi et al., 2023).

In response to these concerns and based on the dynamic capability of AI, this research tries to fill these gaps by applying and examining the potential of ML models to strengthen BI in Muscat's real estate sector. By compiling a real-world dataset of housing prices from various areas in Muscat, we develop and evaluate multiple ML models, including Linear Regression, Ridge Regression, Gradient Boosting Regressor, Random Forest Regressor, and Support Vector Regressor (SVR), to predict property values. We apply hyperparameter tuning to optimize model performance and use grid search to systematically explore a predefined set of hyperparameter values to find the best model performance on the train set. The efficiency of applied models is examined and compared using three performance evaluation metrics, i.e., Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) at the end. Through this comparative analysis, we aim to identify the most effective ML approaches for property valuation in Muscat, thereby demonstrating the practical impact of ML on BI and decision-making in the real estate industry.

2. Theoretical Framework

AI is introduced to be capable of thinking like humans, behaving like humans, thinking rationally, and acting rationally (Russell & Norvig, 2020). Today, AI's subfields are increasingly utilized in operational decision-making and predicting processes, which have obviously improved the marketing efficiency through various kinds of abilities. Analysis and property valuation in real estate fields are extensively revolutionalized using ML and other AI subfields. In ML, plenty of variety of algorithms are offered that are utilized in predictive analysis, including support vector machines, decision trees, and neural networks. Such algorithms are able to identify patterns, learn from patterns, and make accurate predictions in various fields (Dey, 2016). These algorithms effectively work for tasks like classification, regression, and time-series forecasting due to their well-adaptation to handling complex, non-linear relationships within data (Páez, 2023).

Through the incorporation of ML into BI systems, businesses are now able to respond to market dynamics, enhance their performance, and obtain competitive advantages (Hamzehi & Hosseini, 2022). ML algorithms improve the BI systems comprehensively by digitalizing and computerizing data processing, revealing hidden patterns, and producing predictive models that could support data-driven decision-making processes (Lim & Zohren, 2021). By giving businesses access to such productive tools, BI systems now facilitate cross-enterprise analysis and collaboration (Kasemsap, 2016).

Various factors, such as pricing and planning strategies, in the real estate industry are enriched by utilizing BI tools and data-driven approaches. Real estate professionals can predict the demands,

NIPRESS

analyze market trends, and evaluate investment risks with a higher level of accuracy compared to traditional approaches using BI technologies. Applying BI tools and ML algorithms to housing markets has boosted property accuracy and efficiency in decision-making (Trieu, 2017). This improvement is due to the power of ML models in identifying the patterns and correlations in data where traditional methods are not able to simply recognize them (Jain et al., 2019). Therefore, utilizing such technologies could have great efficiency and enhance transparency in businesses like the real estate and housing markets.

While the real estate sector plays a crucial role in economic growth among the Gulf countries and Oman in particular, it faces specific challenges such as inconsistent market transparency and different regulatory frameworks. The need for advanced analytical tools such as BI and ML to cope with these limitations and constraints and to support sustainable urban development is highlighted by recent research (e.g., Alsaidi et al., 2025). While services like e-government and smart city are initiated and construct the foundation of digital transformation in Oman, the application of BI and ML in sectors such as real estate remains in its early stages (Al Abdulsalam et al., 2024). These pieces of literature emphasize the importance of developing ML applications to improve BI practices suited specifically for the Oman real estate sector.

Property prices are influenced by complex interactions of factors, and traditional methods often struggle to have an accurate valuation. By incorporating ML models into the real estate industry, the accuracy of house price estimations has greatly improved. ML algorithms have shown great performance in interpreting these complicated variable relationships. Over the past several decades, researchers have applied ML as an effective approach in order to build predictive models of house prices in various cities. For instance, Manasa et al. (2020) presented a machine learning-based prediction of house prices using regression techniques for Bengaluru city. One study by Jha et al. (2020) used various ML techniques to predict housing prices on a dataset from Florida that achieved high accuracy levels. Gao et al. (2019) introduced a multi-task learning approach to optimize prediction accuracy. This study emphasized the importance of spatial features in real estate valuation. In another study, Imran et al. (2021) provided contextual insights into local market dynamics by examining the application of ML algorithms within the housing market of Islamabad. Research by Gupta and Tham (2019) explored the foundational theories behind AI and ML, which offered critical viewpoints on the technical foundations that support intelligent systems in domains like property valuation. Tarika and Singh (2021) proposed a comprehensive ML-based framework to estimate property values, in which a range of influencing factors are incorporated into an end-to-end prediction pipeline.

While the reviewed literature validates the powerful impact of AI, ML, and BI systems on predictive analysis, specifically in the real estate sector, these technologies have not been explored within Oman's real estate market. In this study, we aim to fill this gap by evaluating multiple ML models, including Linear Regression, Ridge Regression, Gradient Boosting Regressor, Random Forest Regressor, and Support Vector Regressor (SVR), to seek and assess their effectiveness in property value prediction within this specific market context. For this purpose, we first develop a dataset of house prices from the data collected across various areas of Muscat, Oman. The findings are expected to demonstrate the practical impact of AI, ML, and BI approaches in the real estate industry and to show how ML can enhance property valuation accuracy and support predictive analysis in Oman's real estate sector.

3. Methodology

This section provides the details of the data collection and dataset creation procedure, applied ML models that are Linear Regression, Ridge Regression, Gradient Boosting, Random Forest, and Support Vector Regression, and finally, the performance evaluation metrics used in this study.

3.1. Data Collection and Dataset Creation Procedure

The creation of a well-structured and diverse dataset is a foundational step toward developing robust and effective ML models. Creating such a dataset involves several steps, including data collection, preprocessing, and organization. The dataset used in this work is real-world data gathered from various sources, such as real states located in various areas of Muscat, Oman. The required data was collected

IN PRESS

through discussion, interviews, real estate websites and listings, online resources, and surveys. Considered data points include property location, size (in square meters), number of bedrooms and bathrooms, zone/area, condition or quality, building type (flat or villa), number of floors, number of schools per area, number of hospitals per area, the percentage of expatriates in the area, and price. Overall, eleven 11 attributes were defined and labeled while gathering the data, as described in Table 1.

Table 1
Defined Attributes for the Created Dataset of Muscat Houses

No	Attributes	Description
1	BT	Building type: Type 1 indicates flats, and type 2 indicates villas
2	NR	Number of rooms
3	SIZE	The proportion of residential land zoned for lots over 100 sq.
4	Zone	Zones in Muscat were ranked numerically from 1 to 12 based on the available facilities and quality of the area
5	NF	Number of floors
6	NB	Number of bathrooms
7	NSPA	Number of schools per area
8	NHPA	Number of hospitals per area
9	EXP	The percentage of expatriates in the city
10	Quality	Includes the availability of building facilities such as a swimming pool, gym, garden, yard, and Wi-Fi
11	PTRATIO	The ratio of students to teachers in primary and secondary schools in the neighborhood.
12	PRICE	House price

The collected data was in a descriptive and unstructured format; therefore, the research team worked to organize them into a structured dataset. In the next step, data preprocessing was applied in order to fix the incorrect, corrupted, incorrectly formatted, null, or incomplete data within the dataset. This step included handling missing values and outlier removal, i.e., identifying and handling data points that are significantly different from others. The created dataset at the end, shaped as shown in Table 2, included 810 rows of the different house details in various areas of Muscat, 11 columns of effective attributes on prices of each house (shown in Table 1), and one last column which indicated the price of houses in Omani Rial. Table 3 shows the 10 top entries of the created Muscat House Price Dataset.

Table 2
Muscat House Price Dataset- 10 Top Entries

	SIZE	NHPA	NF	NB	NR	NSPA	BT	EXP	PTRATIO	Zone	Quality	PRICE
1	715	4	3	7	4	12	1	12.229	4.5	1	5	740830
2	1752	4	3	6	7	9	2	228.665	13.6	6	5	620000
3	1300	4	2	6	7	10	2	270.567	13.6	3	5	600000
4	600	4	1	3	3	10	2	270.567	13.6	2	5	595000
5	600	4	2	3	3	10	2	270.567	13.6	2	5	590000
6	500	4	1	3	3	10	2	270.567	13.6	2	5	500000
7	500	4	1	2	2	10	2	270.567	13.6	2	5	500000
8	495	4	1	2	3	10	2	270.567	13.6	2	5	495000
9	490	4	1	2	3	10	2	270.567	13.6	2	5	495000
10	500	4	1	2	3	10	2	270.567	13.6	2	5	495000

The dataset was structured in Excel, and its .CSV format was imported into the Jupyter Notebook dashboard for further analysis.

3.2. Applied ML Models

An ML model applies algorithms to learn patterns from data and makes decisions or predictions without explicit programming. In general, ML models are categorized into four different categories based on their learning paradigm as follows (Russell & Norvig, 2020):

IN PRESS

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

Accordingly, a model could be selected based on the problem type and available data. In this study, given the labeled nature of the acquired data, the applicability of supervised ML is warranted for training the model. Typically, the primary objective of a supervised learning algorithm is to determine a function that establishes a relationship between the input variables (x) and the output variable (y). In the context of this particular study, the input variables are the features with the greatest impact on house prices. Identifying these influential features is a crucial process conducted through preprocessing, exploratory data analysis, and visualization, documented in section 4.

Generally, for the supervised learning model, two types of algorithms are proposed: regression and classification (Hastie et al., 2009). Regression models are applicable when the problem is dealing with the prediction of continuous outcomes. They analyze relationships between the target variable and one or more feature variables. Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

In this study, based on the nature of the problem and the type of collected data, various regression models are applied to the collected data for model training. Given the size of the dataset, applying more complex models such as Neural Networks was not feasible. Hence, the selected models to be applied in this study include Linear Regression, Ridge Regression, Gradient Boosting Regression, Random Forest Regression, and Support Vector Regression (SVR).

Linear Regression is a kind of regression model among supervised algorithms that assumes a linear relationship between input (feature variables) and output (target variable) (Montgomery et al., 2021). The algorithm creates a linear equation that best fits the observed data.

Simple Linear Regression models the relationship between one independent variable (X) and a dependent variable (Y). Mathematically, it can be formulated as (Hastie et al., 2009):

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

in which Y is the dependent variable, X is the independent variable, β_0 is the intercept of the regression line, $\beta_1$is the slope of the regression line representing the change in Y for a one-unit change in X, and ε is the error term, accounting for the variability in Y not explained by X.

Multiple Linear Regression is an extended form of simple linear regression that models the relationship between multiple dependent variables and target variables. Mathematically, it can be formulated as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Gradient Boosting Regressor uses an ensemble learning approach where models are built sequentially to correct the mistakes of the previous models, improving overall accuracy. This approach effectively combines multiple weak models to create a stronger predictive model. Considering a given training dataset with input variables of $X = (x_1, x_2, x_3, \dots, x_n)$ and corresponding target variables of $Y = (y_1, y_2, y_3, \dots, y_n)$, the goal is to find a function $F(X)$ that maps input to outputs by minimizing a differentiable loss function $L(y, F(x))$.

Random Forest Regressor combines multiple decision trees to improve predictive accuracy and control overfitting in regression tasks. Each decision tree in the ensemble is built using a random subset of the training data and features, and the final prediction is obtained by averaging the predictions of all individual trees (Ho, 1998).

Support Vector Regressor (SVR) is a regression technique that uses the principles of Support Vector Machines (SVM) to predict continuous outcomes. In SVR, a hyperplane is constructed in some high-dimensional feature space with the attempt to fit data within the user-specified margin of tolerance.

IN PRESS

Ridge Regression is a linear regression that addresses the problem of multicollinearity among predictor variables.

Besides, each model has different hyperparameters, which can influence the performance of the target. In this study, a set of hyperparameters configured, as shown in **Error! Reference source not found.**, is to be examined on the train set.

Table 3
M: Models and Their Fine-Tuning Hyperparameter Sets

Model	Parameters	Range
Linear Regression	fit intercept	True, False
	positive	True, False
Ridge Regression	alpha	0,0.75,1.5,2.25,3,3.75,4.5,5.25,6
	max iter	7500,10000,12500,15000
	solver	auto, svd, cholesky, lsqr, sparse_cg, sag, saga
Gradient Boosting Regressor	loss	squared_error,absolute_error,huber,quantile
	learning rate	0.01,0.1,0.5
	n estimators	1,100,200
	subsample	0.1,0.5,0.75,1
	criterion	friedman_mse, squared_error
	min samples split	2,3,4
Random Forest Regressor	n estimators	50,100,200
	max depth	4,7,10
	Min samples split	2,3
Support Vector Regressor (SVR)	kernel	linear, poly, sigmoid
	degree	1, 2, 3
	C	0.1, 0.5, 1.0, 1.5
	epsilon	0.1, 0.25, 0.4

IN PRESS

3.3. Performance Evaluation Metrics

The data was split into training and testing sets to evaluate the performance of the ML model effectively. Typically, the dataset was split with a ratio of 80:20, where 80% of the data was used for training the models and 20% was reserved for testing. In the train set, we used 10-fold cross-validation.

Each ML model was trained using the training dataset. We employed grid search for hyperparameters required by each model. The performance of each model was evaluated using three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

- Mean Absolute Error (MAE): This metric measures the average magnitude of errors in a set of predictions without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation.

- Mean Squared Error (MSE): MSE is a performance evaluation metric that calculates the average of the squared differences between the observed values in the statistical study and the predicted values from the applied ML model.

- R-squared (R^2 Score): This metric specifies the proportion of variance in the dependent variable that could be explained by the independent variable, namely, a clear view of how much variance is explained by the model. The R^2 score typically ranges from 0 to 1, with higher values indicating a better fit of the data into the regression model.

IN PRESS

4. Results

4.1. Exploratory Data Analysis

A key step after the data collection and creation of the required dataset in ML projects is Exploratory Data Analysis (EDA). EDA helps to understand the relationships between the feature variables and target variable(s) within a dataset. It could provide insights into the structure and patterns in a dataset that helps to decide about the model selection (VanderPlas, 2016).

In this study, we applied EDA to handle missing data, understand the data distribution, and identify the correlations between feature variables and target variables that are reported below. Data analysis and coding for this study were done through Python programming language and in the Jupyter Notebook. Besides, five libraries were used: Pandas to read, analyze, manipulate, and store the dataset, NumPy to work with numerical data and mathematical calculations and functions in Python, Matplotlib for plotting, static and interactive visualizations in Python, Seaborn for making statistical graphics in Python, and Scikit-Learn library, also known as 'sklearn' to implement ML models and statistical modeling. Below are the steps followed for the data analysis in the Jupyter Notebook.

4.1.1. Handling Missing Data

Addressing missing data is essential for ensuring the robustness and reliability of ML models. Recent studies underscore the critical impact of missing data on the performance and generalization capabilities of ML models. Missing values in this study were identified using tools like summary statistics and specialized functions (e.g., Pandas describe() and isnull() in Python). Imputation and interpolation are applied to fill the missing values. Imputation involves filling in missing values with estimated or predicted values. Simple imputation methods include filling missing values with mean, median, or mode of the respective column. Interpolation estimates missing values based on existing values in a sequential dataset.

The dataset had a final checking of the null values. The final checking result is reported below, which shows that the dataset was cleaned from the null values appropriately.

```
SIZE        0
NHPA        0
NF          0
NB          0
NR          0
NSPA        0
BT          0
EXP         0
PTRATIO     0
Zone        0
Quality     0
PRICE       0
dtype: int64
```

Table 4 shows the output of the Pandas describe() function, which is applied to get the summary statistic for handling the missing values. 'Count' is specified as the number of data in a specific column, 'Mean' is the mean value of the data in a specific column, 'Std' is the observation's standard deviation, 'Min' is the Minimum of the values among the data in a specific column, '25%, 50%, and 75%' are the lower, middle and upper percentile, and 'Max' indicates the maximum of the values among the data in a specific column.

Table 4
Statistic Values

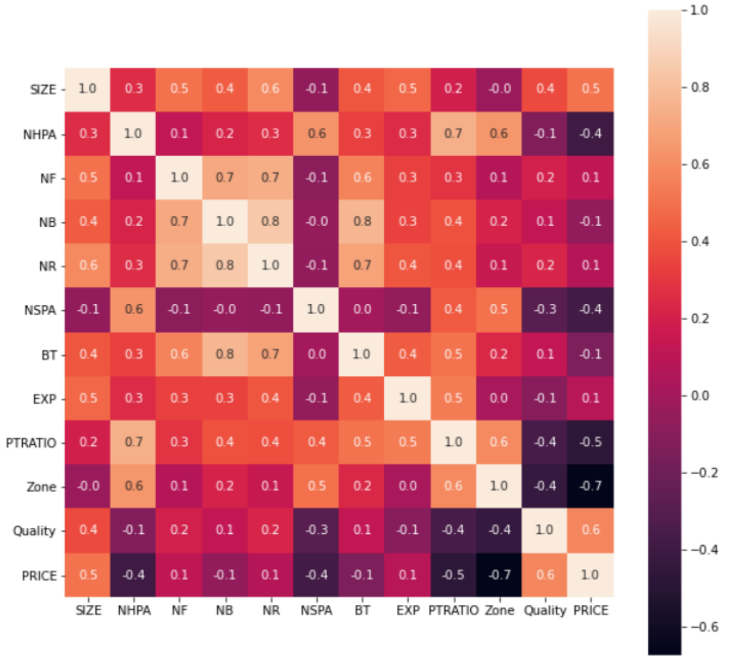
	SIZE	NHPA	NF	NB	NR	NSPA	BT	EXP	PTRATIO	Zone	Quality	PRICE
count	810.00	810.00	810.00	810.00	810.00	810.00	810.00	810.00	810.00	810.00	810.00	810.00
mean	278.33	6.07	1.55	3.40	3.36	12.49	1.30	167.06	24.80	6.06	4.23	166018.23

std	131.09	3.31	0.64	1.55	1.59	3.73	0.46	114.48	19.58	4.27	1.13	117538.42
min	60.00	3.00	1.00	1.00	1.00	1.00	1.00	12.23	4.50	1.00	1.00	22000.00
25%	180.00	4.00	1.00	2.00	2.00	10.00	1.00	12.34	11.60	2.00	4.00	80000.00
50%	310.00	4.00	1.00	3.00	3.00	12.00	1.00	228.67	13.60	6.00	5.00	129000.00
75%	350.00	9.00	2.00	4.75	4.00	13.00	2.00	270.57	50.20	10.00	5.00	230000.00
max	1752.00	14.00	3.00	7.00	7.00	18.00	2.00	291.93	50.60	12.00	5.00	740830.00

4.1.2. Exploratory Data Analysis Using Visualization

In this study, two different plots were used to display the dependency between the feature variables and target values, as the heatmap shown in Figure 1 and the pairplot shown in Figure 2. Seaborn and matplotlib. plot libraries were used to draw these plots in the Jupyter Notebook platform.

Figure 1
Heatmap Plotting of the Correlation Between Features

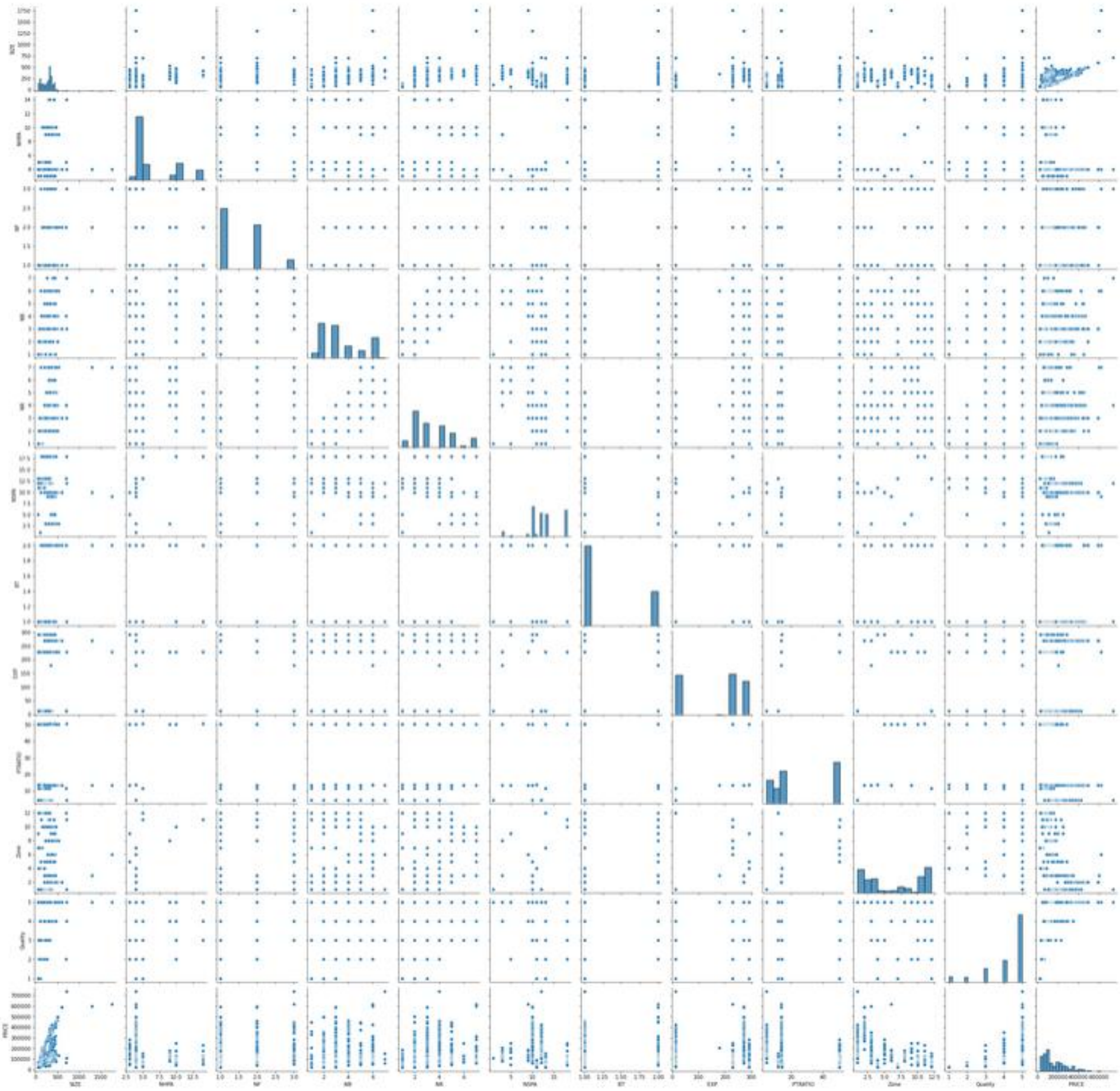


The x-axis and y-axis in the heatmap plot represent the variables. Each cell in the heatmap represents a data value indicating the degree of correlation between pairs of variables. The data values vary in a range between -1 to 1. As much as the absolute value of a cell is closer to 1, the correlation of the corresponding variable is higher. Our target in this specific work is to find the correlation between PRICE and other variables. For instance, the data value shown for PRICE and Zone is '-0.7', indicating a strong dependency between these two variables.

Pairplot shown in Figure 2 depicts the dependency between every pair of attributes. In this study, the target is to figure out the dependencies between each attribute, with PRICE as our target value.

As discussed earlier, data visualization helps in uncovering the correlations between the features and the target variable. Heatmap and pair plots enable us to identify which features have a significant impact on house prices and which are less relevant. This information guides the feature selection process, focusing on the most influential variables and improving the model's predictive power. Data visualization results in the highest dependency of target value (Price) with four feature variables: Size, Zone, Quality, and PTRATIO, which are good candidates for training the ML model in this study. To select the essential features, data analysis continued with deriving the correlation matrix reported below.

Figure 2
Pairplot



4.1.3. Identifying Correlations

Determining the correct input features of the training dataset will provide adequate knowledge to the ML model so that the model can accurately predict the output. Correlation functions and visualization are utilized as the most useful solutions for finding the correct input features. Table 5 is the correlation matrix between features derived by the cormatt() function in the Jupyter Notebook. As can be seen, the essential features are Size, Zone, and Quality, considering a threshold value of 0.5.

Table 5
Correlation Matrix

	SIZE	NHPA	NF	NB	NR	NSPA	BT	EXP	PTRATIO	Zone	Quality	PRICE
SIZE	1.000000	0.256563	0.497060	0.425567	0.593027	-0.061947	0.411711	0.468672	0.190985	-0.028455	0.370628	0.506168
NHPA	0.256563	1.000000	0.122808	0.217713	0.260136	0.629173	0.333700	0.253426	0.741550	0.636956	-0.124897	-0.395744
NF	0.497060	0.122808	1.000000	0.713529	0.736341	-0.088878	0.564305	0.337335	0.286484	0.070735	0.199127	0.115943
NB	0.425567	0.217713	0.713529	1.000000	0.845864	-0.049013	0.768376	0.342778	0.398334	0.211667	0.125499	-0.060091
NR	0.593027	0.260136	0.736341	0.845864	1.000000	-0.060719	0.690961	0.411227	0.380443	0.144994	0.227027	0.067698
NSPA	-0.061947	0.629173	-0.088878	-0.049013	-0.060719	1.000000	0.004704	-0.056698	0.424350	0.497833	-0.284073	-0.378309

BT	0.411711	0.333700	0.564305	0.768376	0.690961	0.004704	1.000000	0.428425	0.505008	0.236217	0.127243	-0.139668
EXP	0.468672	0.253426	0.337335	0.342778	0.411227	-0.056698	0.428425	1.000000	0.537745	0.024871	-0.104422	0.080616
PTRATIO	0.190985	0.741550	0.286484	0.398334	0.380443	0.424350	0.505008	0.537745	1.000000	0.590605	-0.356922	-0.471626
Zone	-0.028455	0.636956	0.070735	0.211667	0.144994	0.497833	0.236217	0.024871	0.590605	1.000000	-0.432907	-0.673073
Quality	0.370628	-0.124897	0.199127	0.125499	0.227027	-0.284073	0.127243	-0.104422	-0.356922	-0.432907	1.000000	0.579962
PRICE	0.506168	-0.395744	0.115943	-0.060091	0.067698	-0.378309	-0.139668	0.080616	-0.471626	-0.673073	0.579962	1.000000

In summary, EDA plays a critical role in the process of creating a predictive model of house prices using ML.

4.2. Feature Selection

In ML and data analysis, feature selection refers to the process of selecting a subset of relevant features from the original dataset while eliminating irrelevant or redundant ones to enhance model performance, reduce computational costs, and improve interpretability.

Identifying the correlations done by heatmap, pairplot, and correlation matrix in the previous step navigates us to a precise and reliable feature selection in this stage. Considering a threshold of 0.5 for the correlation value, Table 6 is the output of the snipped code applied to derive the selected features.

Table 6
Feature Selection Result

	Corr Value
SIZE	0.506168
Zone	-0.673073
Quality	0.579962
PRICE	1.000000

Feature selection helps to eliminate irrelevant or redundant data and enhance the predictive accuracy of the model. Using this, we shaped our new dataset, which contains the most effective feature variables, such as SIZE, Zone, and Quality, and the target value, which is PRICE. Table 7 shows the top five entries of the new dataset after applying the feature selection. Now, our dataset is ready to be given to our ML model for training.

Table 7
New Dataset- Top Five Entries

	SIZE	Zone	Quality	PRICE
1	715	1	5	740830
2	1752	6	5	620000
3	1300	3	5	600000
4	600	2	5	595000
5	600	2	5	590000

4.3. Data Splitting

Generally, data splitting in ML and statistics refers to dividing the dataset into distinct subsets. It ensures that models generalize well and prevent overfitting by providing reliable performance estimates on unseen data (Goodfellow et al., 2016). Data splitting divides the dataset into two or sometimes three subsets as training, testing, and validation (optional) sets. The training set is used to train the model. Using this subset, the model learns patterns, relationships, and parameters. The test set applies to evaluate the final model after training.

Data splitting in this study includes two steps: first, to separate feature values from target values, and next, to form training and testing data. Table 8 is the result of splitting the feature values from the target values, showing the five top entries.

IN PRESS

IN PRESS

Table 8

Data Splitting- Feature values, Top five Entries

	SIZE	Zone	Quality
1	715	1	5
2	1752	6	5
3	1300	3	5
4	600	2	5
5	600	2	5

To have the required testing and training set, the dataset was split with a portion of 80% and 20% for the training set and test set, respectively. In order to ensure that the data is distributed randomly and to have a fair representation in both sets, we used shuffling while splitting. It helps improve the robustness and generalizability of the ML model. The next step is to apply our training models to the training set and test their performance using the test set.

4.4. Training the Models

Generally, training the model refers to the process of learning an ML model to make a decision or prediction based on a given data. For this purpose, the model will be fed with data, its parameters will be adjusted, and its performance will be assessed iteratively until the expected performance is achieved.

In order to have effective training, the dataset must be prepared properly. This preparation includes data cleaning, handling missing values, normalizing numerical values, and splitting the dataset into training, validation (as an optional subset), and test subsets (Goodfellow et al., 2016). The preparation steps that were done on our dataset in the previous steps were the prelude to model training.

Various models could be selected based on the type of studied problem. As discussed earlier, due to the nature of the studied problem and the type of collected data, this study leverages five ML models to train the model for predicting house prices.

- Linear Regression,
- Ridge Regression,
- Gradient Boosting Regressor,
- Random Forest Regressor,
- Support Vector Regressor (SVR)

Furthermore, each model has different hyperparameters that can control the training process and influence the performance of the model. Hyperparameters are set before training and tuned manually or through techniques like grid search or random search. We used grid search to systematically explore a predefined set of hyperparameter values to find the best model performance on the train set. Table 4 shows the set values of hyperparameters that achieved the best performance on the train set. The best values set for the hyperparameters are shown in the last column of Table 9.

Table 9

Machine Learning Models and Their Fine-Tuning Hyperparameter Sets

Model	Parameters	Range	Best
Linear Regression	fit intercept Positive	True, False True, False	True False
Ridge Regression	alpha max iter Solver	0,0.75,1.5,2.25,3,3.75,4.5,5.25,6 7500,10000,12500,15000 auto, svd, cholesky, lsqr, sparse_cg, sag, saga	0 15000 sparse_cg
Gradient Boosting Regressor	loss learning rate n estimators subsample criterion min samples split	squared_error,absolute_error,huber,quantile 0.01,0.1,0.5 1,100,200 0.1,0.5,0.75,1 friedman_mse, squared_error 2,3,4	squared_error 0.5 200 1 squared_error 2

Random Forest Regressor	n estimators max depth Min samples split	50,100,200 4,7,10 2,3	100 7 3
Support Vector Regressor (SVR)	kernel degree C epsilon	linear, poly, sigmoid 1, 2, 3 0.1, 0.5, 1.0, 1.5 0.1, 0.25, 0.4	sigmoid 3 1.5 0.4

4.5. Testing and Evaluation of the Models

Through the previous section, the model is trained on the train set, and the next step here is to test and evaluate the performance of our model. For testing, the models were fed by the test set, and the model's predictions were generated based on the input features. Table 10 demonstrates a sample output of one of the applied models (linear regression). In this table, y_predict indicates the predicted values generated by the model, and y_test indicates the real values in the test dataset.

Table 10
Sample output of Testing Model- Real Values vs. Predictive Values

	y_predict	y_test
1	238364.147590	270000.0
2	106232.220693	95000.0
3	202858.155910	155000.0
4	134369.730505	105000.0
5	142791.298622	180000.0
...	...	...
198	175657.316512	110000.0
199	65694.595997	66500.0
200	241210.142241	180000.0
201	80788.306289	65000.0
202	190447.901866	130000.0

As discussed earlier, in this study, three performance evaluation metrics are used to examine the efficiency and accuracy of the applied predictive ML models: MAE, MSE, and R^2. The applied models were evaluated through these metrics with default hyperparameters (Table 11), and after tuning them, the outcome is shown in Table 12.

Table 11
Models' Performance with Default Hyperparameters

Model	Train			Test		
	MAE	MSE	R2	MAE	MSE	R2
Linear Regression	46801.79	3512389557	68.8	51481.35	4190132723	71.38
Ridge	46801.14	3512463516	68.8	51479.96	4190133403	71.38
Gradient Boosting Regressor	14413.43	630512650	94.4	10692.59	339308956	97.68
Random Forest Regressor	14248.02	697130988	93.81	7619.23	266080933	98.18
SVR	86493.53	12407385893	-10.2	92665.01	16191090927	-10.6

IN PRESS

Table 12

Models' Performance after Fine-tuning Hyperparameters

Model	Train			Test		
	MAE	MSE	R^2	MA	MS	R^2
Linear Regression	51481	4190132723	71.37	46801	3512389557	68.80
Ridge	51481	4190132723	71.37	46801	3512389557	68.80
Gradient Boosting Regressor	6305	200637882	98.62	15203	789648193	92.98
Random Forest Regressor	7798	261208009	98.21	14108	680490822	93.95
SVR	78181	9735515957	33.49	72050	7873143536	30.07

The tuning process highlights the importance of choosing appropriate values for regularization, learning rate, number of estimators, and other model-specific parameters to achieve the best performance. The R² score provides a measure of how well the data fits the applied model. The R² score typically ranges from 0 to 1, with higher values indicating better fit. MAE measures the average magnitude of the errors, and MSE measures the average of the squares of the errors in a set of predictions. Lower MA and MSE values indicate better model performance, as it means that the predictions are closer to the actual values. This analysis suggests that the Random Forest Regressor and Gradient Boosting Regressor provide a good balance between training performance and generalization, making it a robust choice for predicting house prices in this dataset. However, Random forest with slightly higher MAE and MSE compared to Gradient Boosting on the training set generalizes better than the test set, indicating less overfitting.

The high rate of R^2 score for both the Gradient Boosting Regressor and Random Forest Regressor indicates a good predictive power of these models, as almost 93% of the variance in the target variable is explained by the feature variables in these models. These results also specify that the selected feature (i.e., size, zone, quality) used to train our models is relatively effective in predicting the target variable (i.e., house prices), and the model is ready for deployment.

5. Discussion

The findings of this study highlight the potential of ML models in optimizing the predictive accuracy of house price estimation, particularly in real estate markets that are both data-limited and structurally complex, such as Muscat, Oman. Among the evaluated models, i.e., Linear Regression, Ridge Regression, Gradient Boosting Regressor, Random Forest Regressor, and Support Vector Regressor (SVR), the ensemble models, specifically Gradient Boosting and Random Forest, demonstrated better performance in terms of R² scores and lower error rates (MAE and MSE). Our findings align with the outcome of the study done by Antipov and Pokryshevskaya (2012), who showed how ML could improve prediction accuracy in diverse housing markets.

The results of this study are applicable not just to the Omani real estate sector but also could be applied to international markets that are facing similar challenges. ML can offer applicable frameworks that support consistency and reliability in regions where real estate data is inconsistent, or valuation methods lack standardization. ML methods could also assist real estate agents in setting more accurate prices for properties in places where the region's economic and geographical situations may not be fully considered by traditional valuation strategies. Besides, ML-based models are able to enhance transparency by reducing human error in property valuation. These predictive systems help estimate housing values more efficiently and can adapt to changes in the market. The study's findings provide valuable insights for a variety of stakeholders (like developers, agents, buyers, and policymakers) by supporting the adoption of data-driven decision-making in both current and future housing markets.

The predictive power of the models explored in this research can largely be related to an appropriate feature selection and careful hyperparameter tunning. Rather than relying solely on algorithm choice, the study emphasizes the importance of data preparation, variable selection, and parameter optimization. These are steps that enable the models to act across different datasets consistently. This becomes particularly relevant when a specific market context, such as Oman's real estate, is being

IN PRESS

explored, the regions where local market dynamics may differ significantly from global trends. Prior studies, such as the work done by Kou et al. (2014) and Antipov and Pokryshevskaya (2012), have demonstrated that combining domain-specific knowledge with ML would greatly enhance predictive performance. Such approaches not only boost prediction accuracy but also provide actionable tools for agents, developers, and policymakers seeking to navigate complex housing markets.

This research also contributes to academics by applying ensemble learning models to real-world property data. Housing data gathered from various areas of Muscat were processed to shape the structured dataset and build the foundation of the analysis. EDA revealed important location-specific factors that drive price differences. Ensemble models, especially Gradient Boosting and Random Forest, have shown a better performance compared to other approaches. The Gradient Boosting model explained approximately 93% of the variation in house prices, which highlighted its strong performance. Factors such as size, location, and quality are found to be key drivers of property value. Furthermore, the research highlights that AI technologies are not just digitizing traditional tasks but also transforming industry workflows by improving operational efficiency.

Despite these positive outcomes, our research includes several limitations. The dataset was restricted to a specific geographical area, which limited the broader applicability of the models. Also, external influences such as interest rates, government regulations, or macroeconomic trends were not considered, which may affect prediction accuracy in real-world applications. Thus, future studies should expand the geographic area of data collection and also explore additional variables to strengthen the models' generalizability and accuracy. Employing a broader range of ML methods, such as those explored by Eshkiki and Mora (2023), may also help address non-linear data relationships more effectively. Integrating models like decision trees or polynomial regression could further enhance model robustness (Hastie et al., 2009), which supports more comprehensive and flexible house price prediction tools. Future work can improve predictive models by addressing the restrictions and expanding the scope of analysis and model employment.

This study concludes by showing the effectiveness of ML models in predictive data analysis, specifically in the real state sector, and also highlighting the growing role of ML in modern BI. This research provides insights for businesses that aim to use data analytics along with ML models for competitive advantages. In summary, ML's incorporation into property valuation offers transformative opportunities for multiple stakeholders, developers, investors, and policymakers by providing fast, reliable, and intelligent insights. These advantages can greatly support smarter investments and promote enhancement in both emerging and existing markets.

Disclosure Statement

The authors claim no conflict of interest.

Funding

The research received grants from TRC (The Research Council) and the Ministry of Higher Education, Research and Innovation of the Sultanate of Oman.

References

Al Abdulsalam, A. S., Al Hashemi, M. M. A. B., Aleissaee, M. Z. S., Almansoori, A. S. H., Ertek, G., & Labben, T. G. (2024). A novel data analytics methodology for analyzing real estate brokerage markets with case study of Dubai. *Buildings*, *14*(10), Article 3068. https://doi.org/10.3390/buildings14103068

Alsaidi, M. R. S., Al Adwan, M. K., & Al-Mamari, S. N. A. (2025). Real estate registration and its legal effects in Omani Legislation. *International Review of Law*, *14*(1), 163-195. https://doi.org/10.29117/irl.2025.0318

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model interpretation. *Expert Systems with Applications*, *39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

IN PRESS

Dey, A. (2016). Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174-1179.

Eshkiki, H., & Mora, B. (2023). Neighbor migrating generator: Finding the closest possible neighbor with different classes. In B. Muller (Ed.), *Proceedings of AISB Convention 2023 Swansea University* (pp. 79-85). Swansea University.

Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. In T. Li, D. Greenhalgh, & S. J. Fong (Eds.), *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 6–10). Association for Computing Machinery. https://doi.org/10.1145/3195106.3195133

Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, T., & Wu, Z. (2019). *Location-centered house price prediction: A multi-task learning approach* [Preprint]. arXiv. arXiv:1901.01774. https://doi.org/10.48550/arXiv.1901.01774

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gupta, P., & Tham, T. M. (2019). *Fintech: The new DNA of financial services*. De Gruyter.

Hamzehi, M., & Hosseini, S. (2022). Business intelligence using machine learning algorithms. *Multimedia Tools and Applications*, *81*(23), 33233–33251. https://doi.org/10.1007/s11042-022-13132-3

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Huang, C., Li, Z., Chen, F., & Liang, B. (2025). *Multimodal machine learning for real estate appraisal: A comprehensive survey* [Preprint]. arXiv. arXiv:2503.22119. https://doi.org/10.48550/arXiv.2503.22119

Imran, U., Zaman, M., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: The case of Islamabad housing data. *Soft Computing and Machine Intelligence*, *1*(1), 11–22.

Jain, N., Goel, P., Sharma, P., & Deep, V. (2019, March 15). *Prediction of house pricing using machine learning with Python* [Paper presentation]. International Conference on Advances in Engineering Science Management & Technology, Dehradun, India. https://doi.org/10.2139/ssrn.3403964

Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). *Housing market prediction problem using different machine learning algorithms: A case study* [Preprint]. arXiv. arXiv:2006.10092. https://doi.org/10.48550/arXiv.2006.10092

Kasemsap, K. (2016). The fundamentals of business intelligence. *International Journal of Organizational and Collective Intelligence*, *6*(2), 12–25. https://doi.org/10.4018/IJOCI.2016040102

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, *275*, 1–12. https://doi.org/10.1016/j.ins.2014.02.137

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), Article 20200209. https://doi.org/10.1098/rsta.2020.0209

Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine learning-based predicting house prices using regression techniques. In S. V. (Ed.), *The 2nd International Conference on Innovative Mechanisms for Industry Applications* (pp. 624–630). IEEE. https://doi.org/10.1109/ICIMIA48430.2020.9074952

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Páez, A. (2023). Machine learning algorithms for predictive modeling: Analyzing a wide range of machine learning algorithms for predictive modeling tasks, including regression and classification. *Australian Journal of Machine Learning Research & Applications*, *3*(2), 190–198.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

IN PRESS

Tarika, U., & Singh, S. (2021). *Project report on house price prediction (An end-to-end ML project)*. Galgotias University. https://103.47.12.35/bitstream/handle/1/9651/BT3083_RPT%20-%20Amit%20Kumar.pdf?sequence=1&isAllowed=y

Trieu, V. H. (2017). Getting value from business intelligence systems: A review and research agenda. *Decision Support Systems*, *93*, 111–124. https://doi.org/10.1016/j.dss.2016.09.019

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. O'Reilly Media.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. In K. Vohra & D. Mayank (Eds.), *Second International Conference on Inventive Communication and Computational Technologies* (pp. 1936–1939). IEEE. https://doi.org/10.1109/ICICCT.2018.8473231

IN PRESS