DOI: 10.24996/ijs.2024.65.3.40

ISSN: 0067-2904

Predicting Real Estate Prices Using Machine Learning in Abu Dhabi

Fatima Isameel Al Marzooqi[1], Abdesselam Redouane[2*]

[1]*Abu Dhabi City Municipality, Abu Dhabi, UAE*

[2]*E-Business, Canadian University Dubai, Dubai, UAE*

Received: 8/10/2022 Accepted: 6/4/2023 Published: 30/3/2024

Abstract

Traditionally, real estate prices were determined based on demand and supply. As the real estate market was unregulated and underdeveloped, brokers and real estate builders had an upper hand in determining the unit prices of residential houses in Abu Dhabi. A pricing gap was eventually noticed. This was a challenge. There is a delay in updating the real estate websites and portal information. Therefore, the need for accurate forecasting of prices has become urgent. With a variety of use case scenarios for machine learning concepts, this paper is dedicated to using the concepts of machine learning to predict the real estate prices of Aldar in the Abu Dhabi region, which comprises 511 residential units, 15 retail shops, and one community center. Decision tree, random forest, support vector machines, and K-nearest neighbors (KNN) algorithms were used to identify which one is better for forecasting these real estate prices. Comparing the generated models, the random forest is the best-performing model, followed by support vector regression, and the decision tree model is the least-performing model.

Keywords: Machine learning; Decision tree; Randomforest; Support vector machine; K-nearest-neighbors.

1. Introduction

With new infrastructure and developmental projects going on throughout the year, the UAE region has seen a major boom in the last decade, and it continues to grow even today. Being a region rich in oil exploration and oil reserves until the last decade, a major structural shift to promote and develop it into a non-oil economic powerhouse and boost the real estate industry and the tourism sector has been the cornerstone of the UAE government's success.

Traditionally, real estate prices were determined based on demand and supply. As the real estate market was unregulated and underdeveloped, brokers and real estate builders had an upper hand in determining the unit prices of residential houses in Abu Dhabi. There is a growing need to predict the real estate prices of Aldar in Abu Dhabi since unreliable real estate websites delay updating real estate information. Therefore, the need for accurate forecasting of prices has become urgent. As the fields of data science (DS) [1], artificial intelligence (AI) [2], and machine learning (ML) [3] rose to prominence, they offered data scientists and researchers an opportunity to implement the concepts of machine learning to better analyze the prices of real estate transactions and also forecast the future prices. As sophisticated machine learning algorithms are designed to forecast the real estate price, it shall bring a sense of calmness and stability to this sector.

[*] Email: abdesselam.r@gmail.com

Abu Dhabi real estate buyers, or home buyers, have been facing the problem of exceedingly high brokerage charges by the real estate agents and the lack of reliable information that would ideally form a basis for price comparisons. The lack of trustworthy real estate websites and databases is a challenge that its residents are facing as of now. We can question the authenticity of these databases all the time. At times, to get a better real estate deal, builders or developers tend to inflate or overhype the valuations of real estate projects. Factors like incomplete data, unreliable data, a lack of authenticity, and flawed valuation lead to incorrect market determination of the prices of residential units in Abu Dhabi. This motivates us to take up this research, which will determine the fair valuation of residential and real estate projects in Abu Dhabi, free from any bias.

The real estate market in Abu Dhabi in 2020 will be competitive despite the aftermath of the pandemic. Both the rental and sales markets are hotspots as compared to other countries in the GCC region. The prime locations of Abu Dhabi, like AI Reem Island, AI Raha Beach, and Saadiyat, are still preferred by investors and tenants.

The paper is structured as follows: Section 2 discusses related work, while Section 3 describes the research methodology. A discussion of the findings is provided in Section 4, while the conclusion is presented in Section 5.

2. Related Work

The housing market and the real estate sector are prone to fluctuations and sudden economic jerks, and the housing industry is highly correlated with several known and unknown factors. [4] performed a study on the impact of the global financial crisis on UAE real estate infrastructure and the construction sector. The literature found evidence of a decrease in profitability of real estate properties resulting from fluctuation in rental prices as well as oversupply due to increased construction. Thus, we would expect house prices to decline gradually after the 2007 economic crisis due to its negative impact.

Baldominos et al. [5] performed a study on identifying real estate opportunities with the help of machine learning. The authors argued that the traditional regression-based model relied heavily on statistics, inputs from annual reports, and data from the financial statement, which are historical data. As these models rely heavily on past data to predict the future, they do not always work in all market conditions. The issue to be solved is that the online listing of house prices is not updated when there are changes in the prices. Hence, there is an opportunity for investors in this market. The authors used a regression approach to build this application. They implemented different models, including regression trees, K-nearest neighbors, support vector machines, and neural networks. Ensemble regression trees performed better than the other approaches.

[6] applied multiple machine learning algorithms to predict real estate prices based on geographical location (longitude and latitude), house age, the total number of bedrooms, number of bathrooms, population, ocean/sea proximity, median income, and house size. The methods used include random forests, multiple regression, support vector machines, gradient-boosted trees, and multi-layer perceptrons (neural networks). Random forest had the highest performance (RMSE = 0.012), followed by bagged regression, which yields RMSE equal to 0.563, and multiple linear regression had the lowest performance (RMSE = 0.70). The study used data collected globally, which might not represent the situation in Abu Dhabi. For this reason, there is still a need for a study targeting our study location.

Park and Bae [7] used a classification approach rather than regression to determine whether the closing price was closer or not to the listing price. Among the features selected were the number of bedrooms and bathrooms, the number of fireplaces, the total area, the cooling and heating systems, the type of parking, etc. They used different algorithms like decision trees, RIPPERS, Naïve Bayes, and AdaBoost. RIPPERS outperformed the other models.

Truong et al. [8] used random forest, XGBoost, and lightGBM. In addition, they applied hybrid regression and stacked generalized regression. These methods have their own pros and cons. For instance, random forest has a lower error rate on the training set but has a problem of overfitting and takes much longer to generate the final model. The hybrid regression method performed better than random forest, XGBoost and lightGBM. Stacked generalization regression performed better in terms of accuracy.

Instead of using the House Price Index (HPI) to estimate the inconsistencies of house prices, [9] applied random forest as a classification method as they were looking for price variance rather than a fixed value. It used the Boston housing dataset from the UCI repository. The authors claim that the model has an acceptable predicted value with an error margin of ±5.

Zaman et al. [10] used many different machine learning algorithms, like linear regression, support vector regression (SVR), random forest, and others. They applied these algorithms to a Pakistani data set, and the most performant model was SVR compared to the rest. They used different metrics for this evaluation, like MAPE, MAE, and RMSE. The dataset consists of over a year's worth of data collected from online stores.

3. Methodology

The following workflow illustrates the steps taken to conduct this research.

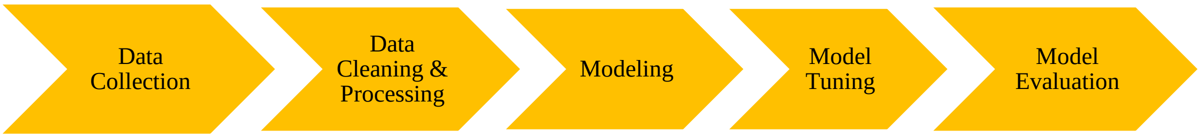

Figure 1: Methodology Phases

In the following sections, these phases are described in detail.

3.1 Data Collection

Data collection was made from primary data and several secondary data sources. The prices of 511 residential house properties were randomly sampled, and property features and historical information were recorded to act as candidate variables that can be used to predict house prices.

Description of the dataset

Table 1: Variables description

Variable	*Description*	*Type*
Community	Community in which the house is located	Categorical
Neighborhood Name	Neighborhood in which the house is located	Categorical
Project Building	project Building under which the house was build	Categorical
Street	Street in which the house is located	Categorical
Floor	The floor in which the house Unit is located	numeric
Unit Number	Unique House ID	numeric
Property Usage	Property Usage (Residential/Commercial)	Categorical
Unit Model	House unit Model(1BHK) =1000 square feet, (2BHK) =1500 square feet(3BHK) =2300 square feet,(5BHK)= 5900 square feet, 590 = (studio apartment with balcony).	Categorical
Unit Sea View	Unit Sea View (none, partial or full view)	Categorical
Unit Marina View	Unit Marina View ((none, partial or full view)	Categorical
Community Facility	Community Facility available on this house complex	plain text
Service Charge Per Year / m2	yearly Service Charge Per square meter	numeric
Private Marina	Availability of private marine	Categorical
Bedrooms	Number of Bedrooms	numeric
Maid Room	Availability of maid room within the house unit	Categorical
No. of Bathroom	Number of Bathrooms	numeric
No of Kitchens	Number of Kitchens	numeric
Car Parks	Number of car parks	numeric
Car Park Size m2	Size of car parks in square meters	numeric
Total Parking size m2	Total parking size in square meters	numeric
Total Unit Measured Area m2	Total house area in square meters	numeric
Terrace/ Balcony Area m2	Total Balcony /Terrace in square meters	numeric
Total Internal Area m2	Total internal area in square meters	numeric
Unit Type	Unit type (Apartment, Bungalow etc.)	Categorical
Year of Construction	Year in which the house was constructed	numeric
Year of Occupancy	Year in which the house was first occupied	numeric
Expected Market Lease Price	Expected Market Lease Price	numeric
Original Selling Price	Original price in which the house	numeric
Price 2015-2020	Price value for the years 2015-2020	numeric

3.2 Data Cleaning and Processing

The variables "community," "neighborhood,” and "street" are the same for all sampled properties. Similarly, property usage and community facility use are constants throughout. Such variables are not useful because they cannot explain any amount of variation in price. Hence, we dropped them from the analysis.

Conversely, a unit number is an ID unique to each property. It does not explain any useful pattern in price because it was assigned for the purpose of identification and does not describe anything about the property. Variable floor, which is supposed to be numeric, has "G," denoting ground floor, and has been changed to "0," and the same was done for "studio," denoting that the property is a studio that cannot have bedrooms/bathrooms. There are no missing values in the dataset.

3.3 Analysis Methods

3.3.1 Decision Trees

Decision trees are a popular machine learning method used for both classification and regression [11], [12]. One of the most common types of decision trees is the recursive partitioning tree (RPART) [13]. For a decision tree algorithm to be used for regression, the information gain should be replaced by a standard deviation reduction. Standard deviation is used to calculate the homogeneity between a sample of numeric values. It is given by the following formula:

$$S = \sqrt{\frac{(x - \bar{x})^2}{n}} \tag{1}$$

where x is the value of a certain attribute; $\bar{x}$ is the mean of these values; and n is the number of these values.

The standard deviation for two attributes is given below [14]:

$$S(T,X) = \sum_{c \in X} P(c)S(c) \tag{2}$$

Where T is the target data set; X is the attribute in question; c is the value that attribute X can take; P is the probability; and S is the standard deviation.
The algorithm uses the standard deviation reduction to decide which attribute to use for branching [14]. The higher the standard deviation reduction, the better. The standard deviation reduction formula is given below:

$$SDR(T,X) = S(T) - S(T,X) \tag{3}$$

For the regression cases, it uses binary partitioning to put the training data into groups such that the overall residual sum of squares is minimized. The prediction rules are saved for use with future datasets.

The algorithm can result in a very deep-rooted and complex tree if control is not exercised. This happens if the model is allowed to emphasize small differences such that some splits result in negligible improvements in the residual sum of squares. For this reason, some control hyperparameters exist that need to be chosen. Some parameters include:

- The minimum number of cases that must exist in a node before a split is attempted (minsplit) E.g., minsplit = 20 would mean that nodes with fewer than 20 observations should not be divided.
- The minimum number of cases that must exist on the resulting terminal not (minbucket) E.g., minbucket = 7 would mean that even if a node satisfies the minimum split, it should not result in a node with fewer than 7 observations.
- Maximum number of nodes to allow (maxdepth).
- complexity parameter (cp) Splits that do not improve the overall lack of fit by the factor of cp are not included.

3.3.2 Random Forrest (FR)

Unlike decision trees, which are based on a single tree, the random forest algorithm uses a "forest" of decision trees; the trees are trained using the bagging technique, which involves fitting decision trees on bootstrap samples [15, 16]. The algorithm samples both observations' attributes to create what is popularly known as "ensembles." During prediction, each tree returns a prediction for each new variable, and the average of the predictions forms the overall predicted value for that case.

As an advanced decision tree, the model has additional hyperparameters that control the sampling process. They include:
ntree – it specifies the number of trees to use.
mtry - specifies the number of features to select in each split.
Smaller values for these parameters may result in a computationally intensive process.

3.3.3 Support Vector Regression (SVR)

Given a training sample, support vector regression approximates a mapping of the provided feature values to a domain of real numbers [17], [18], [19]. [19]. SVR is a generalization of the famous SVM classifier developed by [17]. The generalization is achieved by the introduction of the e-insensitive area around the function to be constructed to perform the prediction. Instead of finding a hyperplane as in SVM, SVR introduces an e-insensitive loss function, i.e., the predicted value in the trained set is less than epsilon from the actual value. The mappings can be linear or nonlinear, kernel tricks are useful ways of adding nonlinearity to the model. The linear case to separate the data can be formulated as follows:

$$y = f(x) = < w, x > + b \tag{4}$$

where $y, b \in R$; $w, x \in R^n$; <. , . > is the dot product.

The figure below illustrates this linear case:

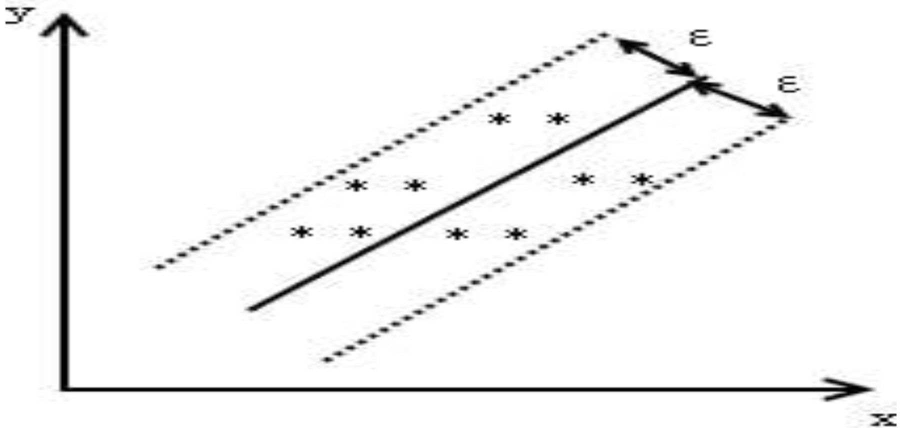

Figure 2: Linear SVR

The objective is to discover f(x) that has at most an epsilon deviation from the actual target value for every training data point and is at the same time as flat as possible [18]. To achieve this, we minimize the norm $||w||^2$. This can be written as a convex problem:

$$Min\ \frac{1}{2}\ \left|\left|w\right|\right|^2 \tag{5}$$

subject to the following constraints:

$$\begin{gathered} y_i - < w, x_i > \ -b\ \le\ \varepsilon \\ < w, x_i > \ +b -\ y_i\ \le\ \varepsilon \end{gathered}$$

The convex optimization problem is feasible, i.e., we can find f(x). However, it is possible to allow for some errors, i.e., to accept data that is outside the tube as defined. We can introduce slack variables, ξi, ξ*, to deal with the infeasibility problem. Figure 3 illustrates the linear SVR with slack variables.

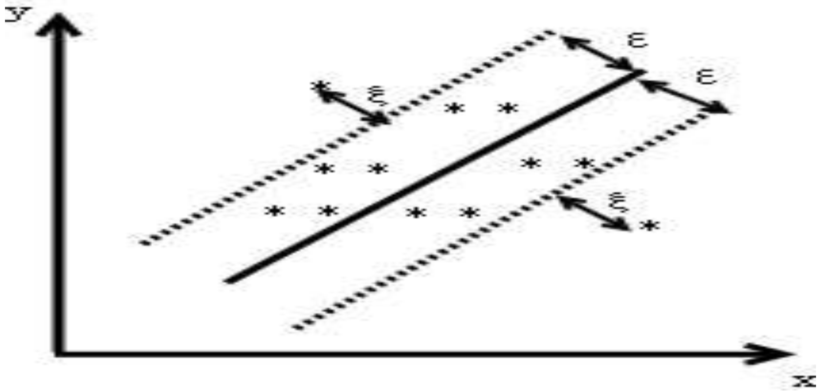

Figure 3: Linear SVR with slack variables

The optimization problem then becomes:

$$Min\ \frac{1}{2}\ \left|\left|W\right|\right|^2 +\ C\sum_{i=1}^{n}(\xi_i +\ \xi_i^*) \tag{6}$$

subject to the following constraints:

$$\begin{cases} y_i - < w, x_i > \ - b\ \ \le\ \varepsilon +\ \xi_i \\ < w, x_i > + b -\ y_i\ \le\ \varepsilon +\ \xi_i^* \\ \quad \xi_i,\ \xi_i^* \qquad\qquad \ge 0 \end{cases}$$

The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. The dual form of the optimization problem in (6) can be solved in its dual form using Lagrange multipliers.

3.3.4 K-Nearest Neighbor (KNN)

The KNN algorithm performs classification based on the features of the k neighboring data points [20]. It is a supervised learning method; given the training data, the algorithm first calculates the distance metric (e.g., Euclidian distance, Manhattan distance, etc.) between pairs of data points, and neighborhoods are then determined based on these distances. After determining the neighbors, the values for the target variable are averaged (only for the neighboring cases), and this is the overall predicted value. The Euclidian distance and the Manhattan distance are given by the following formulas:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (7)$$

$$\text{Manhattan Distance} = \sum_{i=1}^{k}|x_i - y_i| \quad (8)$$

where x_i; y_i are different values for the different attributes being evaluated.

3.4 Model Tuning

In all the above models we mentioned, hyperparameters needed to be chosen to maximize model performance. A common method of finding such values is by iterative search, where a model is fitted each time with different parameter specifications until a combination of parameters that minimizes error is reached. In this study, models will be tuned using the caret package in R, and then the best model will be chosen based on root mean squared error.

3.5 Performance Measures

Error measures such as root mean squared error (RMSE), R squared (R squared), and mean absolute error (MAE) and their standard deviations will be computed. RMSE will be used to make comparisons between model performances. RMSE, MAE, and R squared are given by the following formulas:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \quad (9)$$

$$MAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n} \quad (10)$$

$$R^2 = 1 - \frac{SSR}{SST} \quad (11)$$

where a is the actual target; p is the predicted target.
SSR is the sum of the residuals squared; SST is the total sum of squares.

4. Results and Data Analysis

4.1 Exploratory Data Analysis

Exploratory data analysis was first done by way of graphical visualization and summary statistics. The aim is to get an overview of how house prices are distributed over time as well as relationships between house features and prices.

For the five-year period between 2015 and 2020, Project al Manara resulted in the most expensive houses; the average price is AED 6,853,955 with a standard deviation of AED 572,465.2. The median price is slightly higher than the mean, which implies that most of the houses on this project had prices higher than the mean and only a few underpriced houses are pulling the mean price down. The skewness coefficient lies between -1 and 1, indicating that the mean is a good representative of typical house prices for the project. Al-Naseem-C had the second-most expensive houses over the period, with a mean price of AED 2,126,322 and a standard deviation of 1,164,236. Here, prices are right-skewed, which means most houses have lower prices with only a few overpriced houses. Al-Barza project houses had the lowest price.

Based on the unit model, 4BHK houses had the highest prices on average (AED 5,643,067), while studios had the lowest prices; this is an implication that house size (square footage) influences pricing. Houses with full sea views were found to have higher prices on average compared to houses with partial sea views (median price = AED 1,876,705). For a marina view, houses with a marina were more expensive (median price = 1,822,006) compared to houses with a partial view. The same case was observed at a private marina. Penthouses and houses with maid rooms were also more expensive than the rest of the houses. There is a declining trend in average house prices for the years 2015–2020; average prices fell from AED 2.14 million in 2015 to AED 1.78 million by 2020.

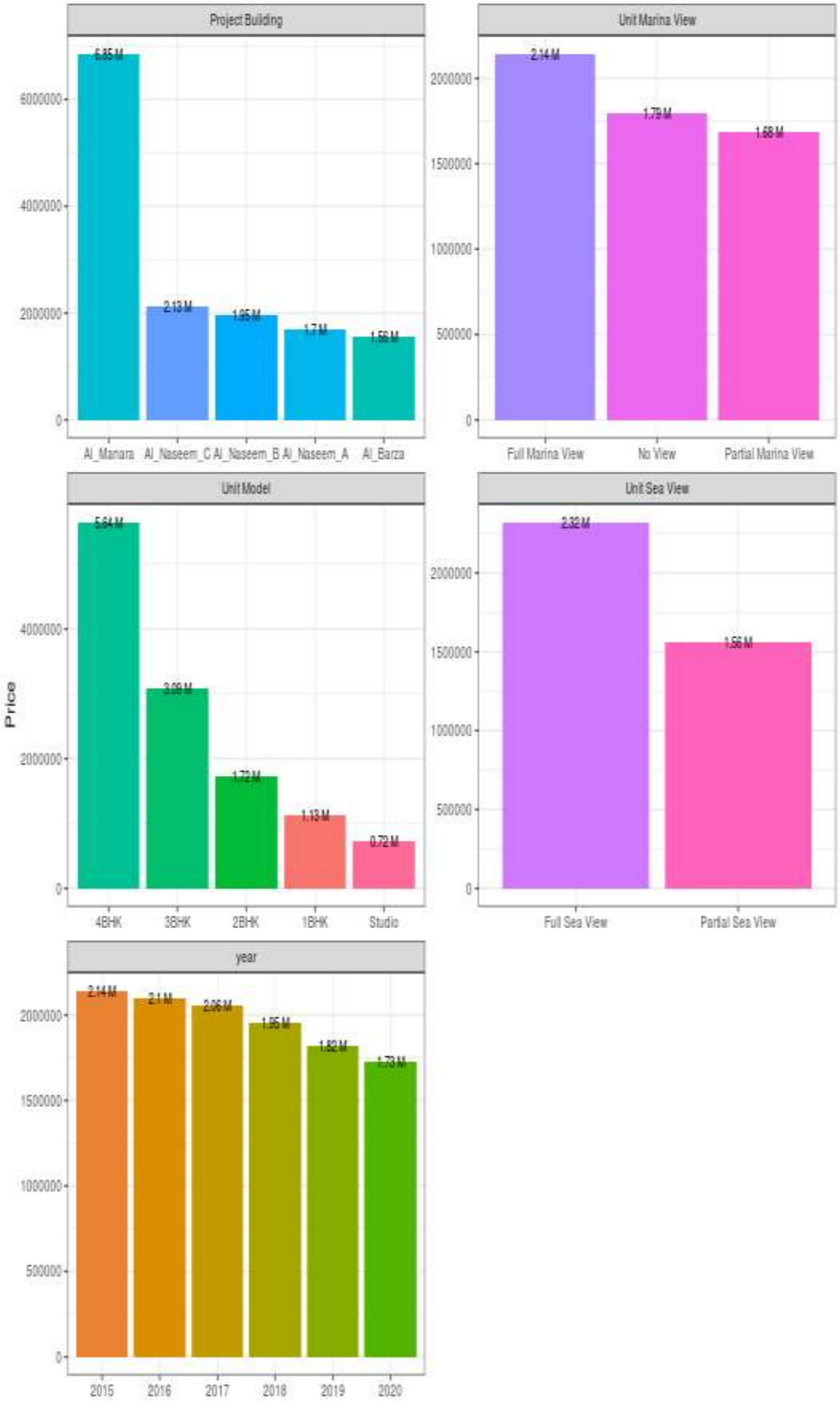

Figure 4: Price per project

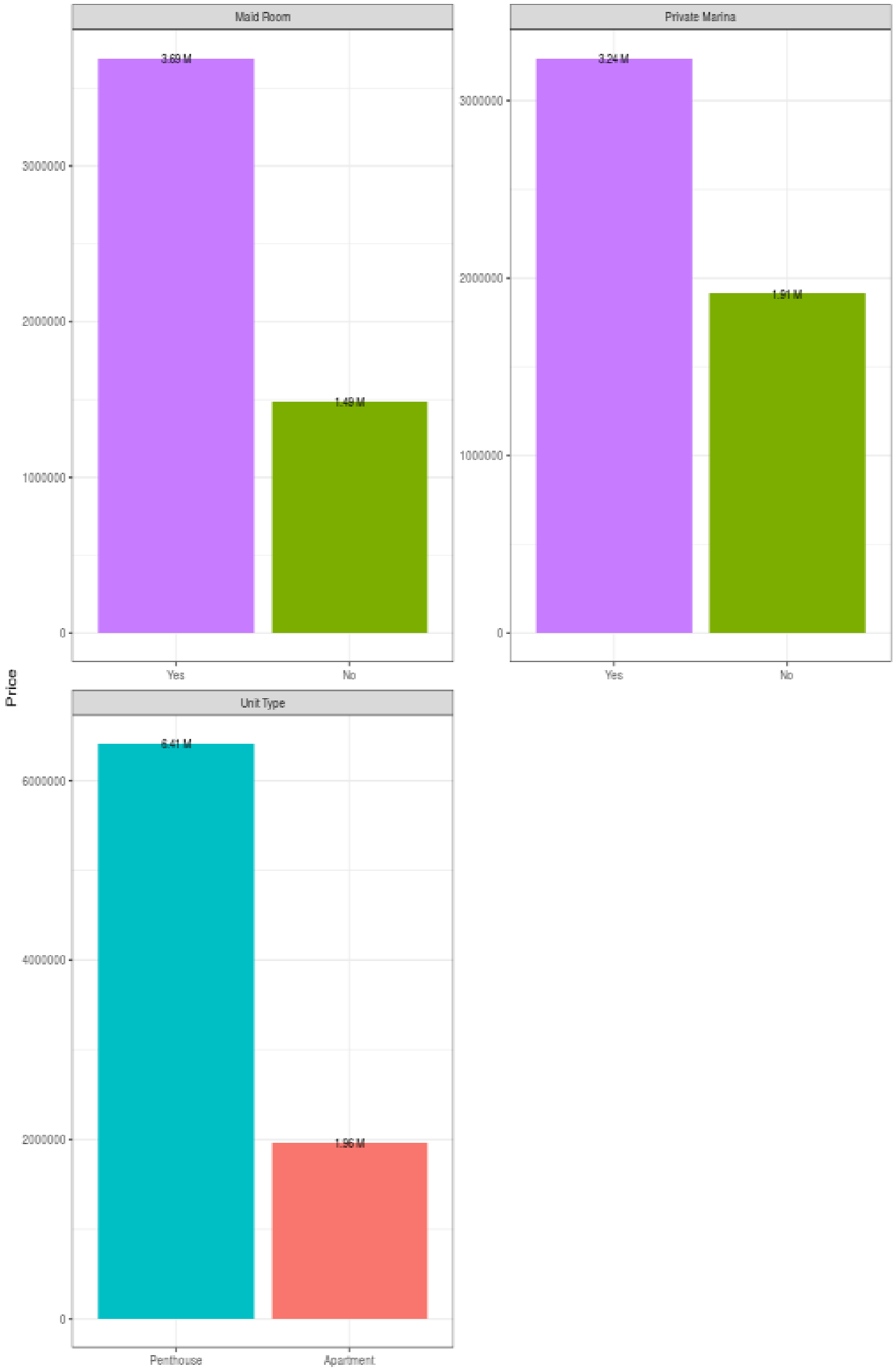

Figure 5: Price per room, ownership, and unit type

From a correlation plot, we found a strong positive correlation between price and all the variables except the floor, year of occupancy, expected lease price, and the current year. In other words, price changes per year are too small to provide evidence of a correlation.

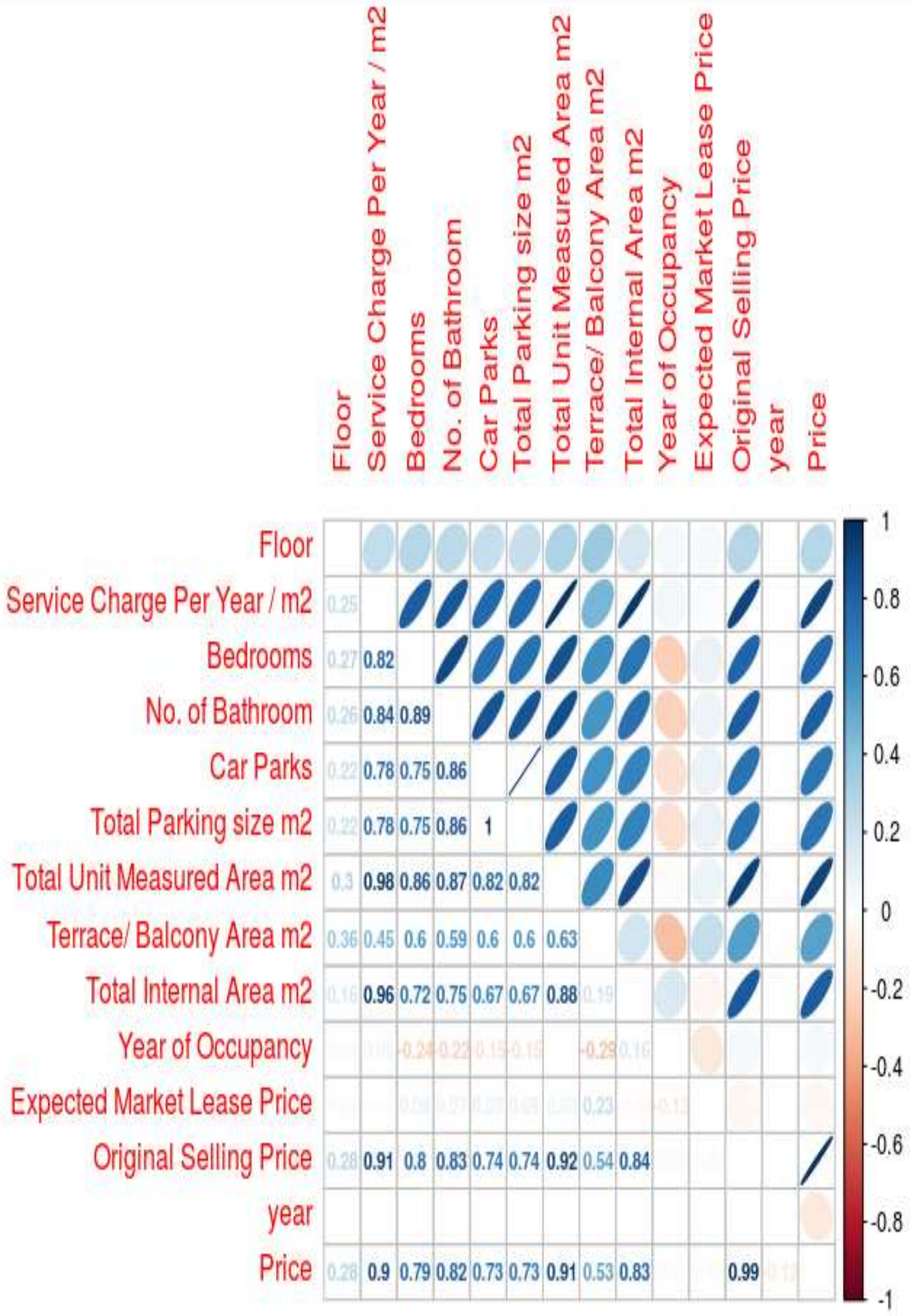

Figure 6: Correlation plot

4.2 Modelling

4.2.1 Decision Tree

As discussed in the previous section, the choice of hyperparameters is important and can affect the performance of the models. Several values of these parameters were tried to come up with a combination of values that result in optimal performance. A model with a complexity parameter of 0.003 and minbucket = 7 and minsplit = 20 was found to be optimal; it yields an RMSE of 233,501. The R-squared of the model is 96.5%. The table below summarizes the findings obtained.

Table 1: Error measures

cp	*RMSE*	*Rsquared*	*MAE*	*RMSESD*	*RsquaredSD*	*MAESD*
0.003	233,501	0.965	179,303.0	12,440.10	0.005	7,576.96
0.003	236,561	0.964	181,080.9	13,575.64	0.006	8,299.24
0.004	254,593	0.959	188,845.0	17,458.73	0.007	9,642.76
0.008	281,723	0.95	200,400.1	18,499.12	0.009	10,256.38
0.01	301,168	0.942	216,252.0	23,962.33	0.012	18,586.67
0.012	329,144	0.931	238,918.4	25,004.91	0.014	13,954.78
0.025	362,290	0.916	266,318.1	36,439.75	0.019	29,683.99
0.107	461,388	0.86	361,578.3	89,348.84	0.058	84,131.19
0.138	654,075	0.72	513,266.7	80,590.02	0.08	45,272.59
0.66	945,155	0.643	677,022.6	293,952.80	0.02	165,954.50

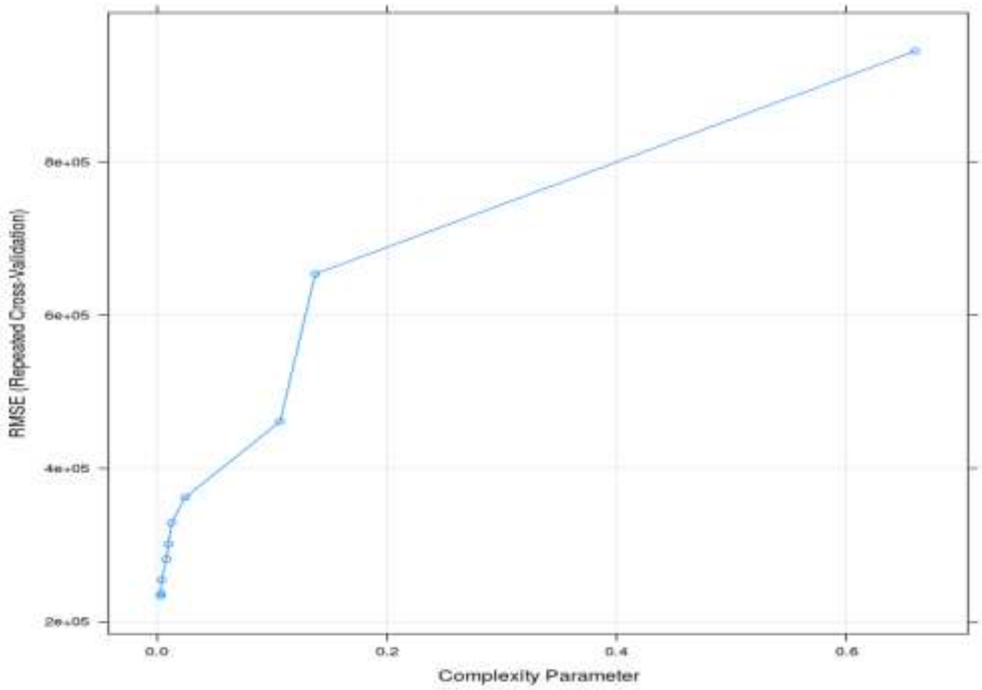

Figure 4: RMSE for different values of CP as in Table 1 above

4.2.1.1 Variable Importance

The objective is to recursively split nodes in a manner that would result in a decrease in the residual sum of squares. The significance of a variable is thus determined by how much the overall RSS decreases when the variable is included in the model. For this case, the variable "original selling price" results in 4.9 decreases in RSS, which is the highest. This is the most important variable; total unit area is the second most important variable; its presence decreases RSS by 3.06. The figure below uses a decision tree to illustrate the variable importance.

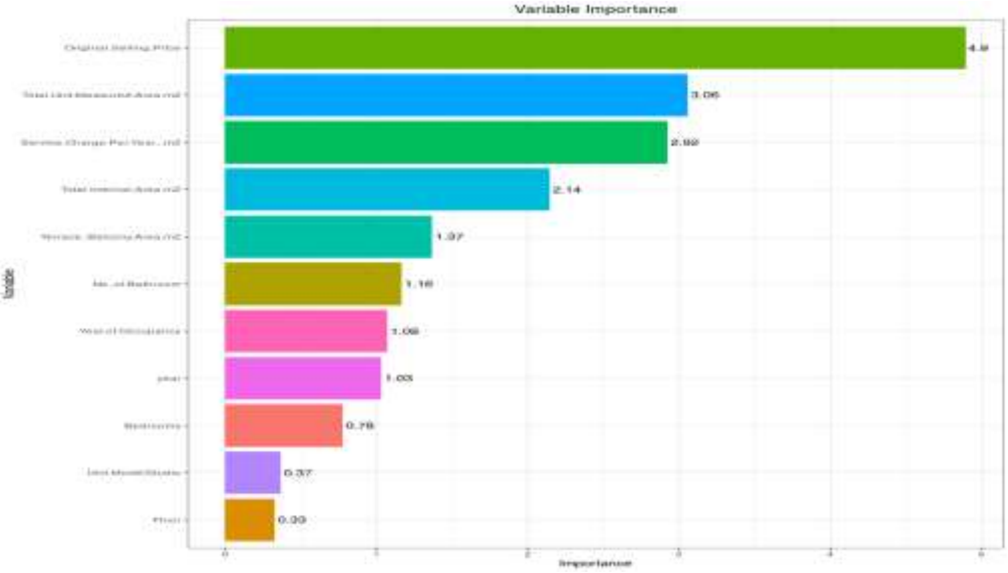

Figure 7: Variable importance

4.2.2 Random Forrest

A random forest model was run with the tuning of parameter mtry. One challenge in random forest tuning is that, unlike mtry (the number of features to sample), which ranges between 1 and the number of variables used, ntree (the number of trees to use) has no limit and can range from 0 to infinity. For this reason, Caret tunes only mtry. Sampling 27 features each time was found to be optimal, with an RMSE of 350.32, while the expected amount of variation explained in price is 99.9%. Table 2 below summarizes the findings.

Table 2: Error measures

Mtry	*RMSE*	*Rsquared*	*MAE*	*RMSESD*	*RsquaredSD*	*MAESD*
2	264154	0.960473	195158	9732.255	0.003852	6018.76
14	55533.05	0.998066	29834.67	5910.374	0.000385	1791.872
27	35032.39	0.999223	13577.85	5402.188	0.000229	1211.264

4.2.2.1 Variable Importance

The original selling price was found to decrease RSS the most, followed this time by the year, service charge, and house size. Unit type and Marina View had the least influence. Variable importance is illustrated in the figure below.

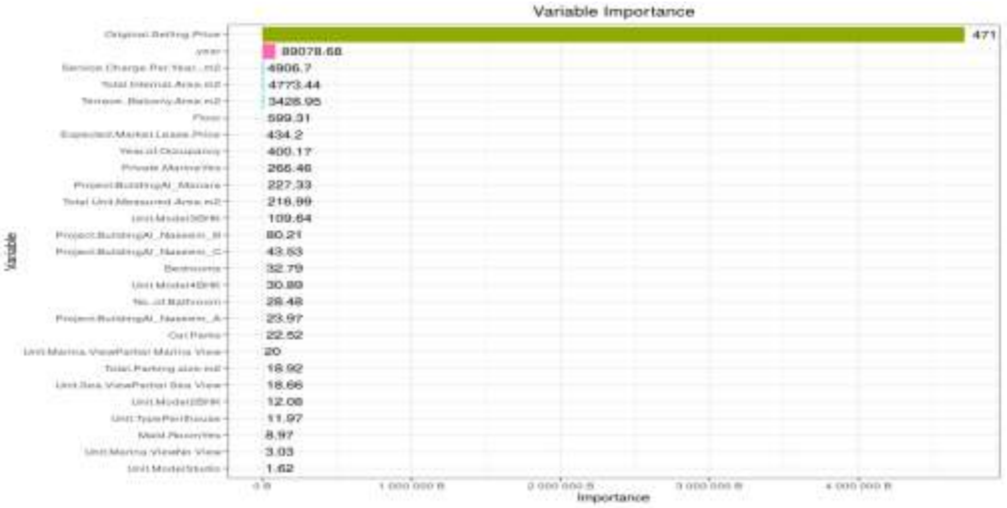

Figure 8: Variable importance

4.2.3 Support Vector Machine

A support vector machine model was fitted using a Gaussian radial basis kernel function; the model was first tuned to optimize the choice of parameters. The optimal choice had epsilon = 0.1, a cost parameter of 32, a sigma of 0.039, and 217 support vectors. The cross-validation error is 77637.01. This is less than what we had with the decision tree. The expected amount of variance in house prices explained by this model is 99.6%. The table below summarizes the findings.

Table 3: Error measures

sigma	*C*	*RMSE*	*R squared*	*MAE*	*RMSESD*	*RsquaredSD*	*MAESD*
0.04	0.25	152,817.70	0.986	94,210.33	44,532.75	0.007	7,504.10
0.04	0.5	105,724.90	0.993	77,570.63	22,405.05	0.002	5,067.42
0.04	1	86,206.12	0.996	69,923.73	8,774.50	0.001	3,424.79
0.04	2	79,715.88	0.996	67,176.62	3,433.48	0.001	2,609.37
0.04	4	78,034.07	0.996	66,522.99	2,612.24	0.001	2,366.88
0.04	8	77,762.67	0.996	66,510.51	2,530.35	0.001	2,256.88
0.04	16	77,651.28	0.996	66,464.71	2,503.86	0.001	2,244.10
0.04	32	77,637.01	0.996	66,454.17	2,504.52	0.001	2,243.99
0.04	64	77,637.01	0.996	66,454.17	2,504.52	0.001	2,243.99
0.04	128	77,637.01	0.996	66,454.17	2,504.52	0.001	2,243.99

4.2.4 K-Nearest Neighbor

A 7-nearest neighbor model was found to be optimal; it led to a cross-validation error equal to 181,671.90; the expected amount of variation in house prices explained by the model is 97.90%, which is slightly lower than what was explained by SVR. The table below summarizes the values of the metrics obtained with different values of K.

Table 4: Error measures

k	***RMSE***	***Rsquared***	***MAE***	***RMSESD***	***RsquaredSD***	***MAESD***
5	187,516.60	0.977	142,449.30	7,607.86	0.004	5,113.54
7	181,671.90	0.979	138,933.70	7,994.46	0.003	5,086.35
9	184,363.80	0.978	140,999.70	8,071.60	0.003	5,056.41
11	182,181.40	0.979	139,506.00	8,039.58	0.003	4,757.25
13	182,589.80	0.979	139,080.30	8,410.12	0.003	4,668.35
15	183,290.40	0.979	139,230.70	8,605.72	0.003	4,727.36
17	183,567.60	0.978	138,783.90	8,401.86	0.003	4,908.19
19	184,230.70	0.978	138,850.90	8,668.13	0.004	5,011.01
21	185,244.40	0.978	138,984.60	9,152.81	0.004	5,197.29
23	185,165.40	0.978	138,644.20	9,664.04	0.004	5,182.65

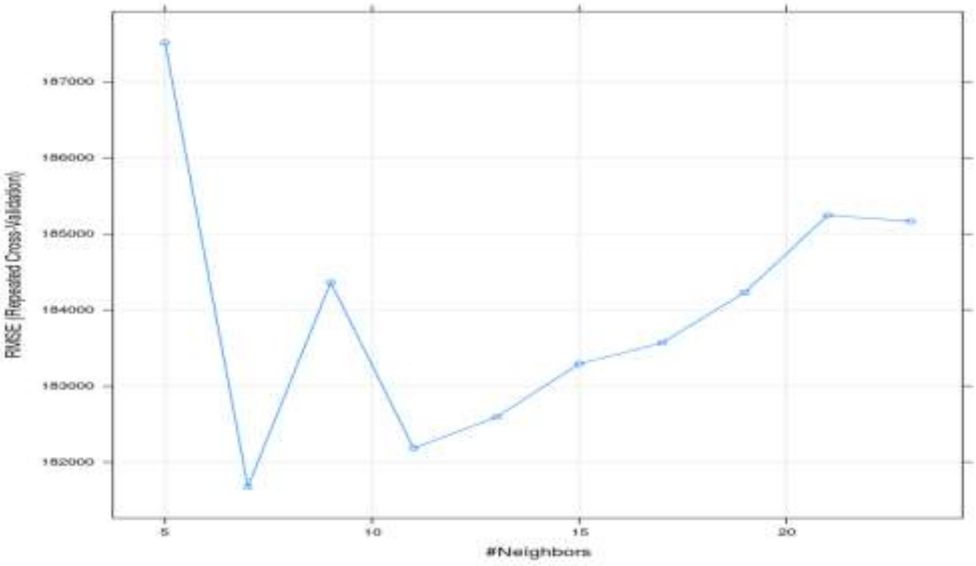

Figure 9: Error measures with different neighbors

4.3 Comparison Between Models and Discussion

Comparing the fitted models, the random forest was found to be the best-performing model, followed by support vector regression, and the decision tree was the least performing. This conclusion was reached based on the cross-validation errors produced. Table 5 below summarizes the findings.

Table 5: Comparison between models with respect to error measures

Model	*RMSE*	*Rsquared*	*MAE*	*RMSESD*	*RsquaredSD*	*MAESD*
Random forest	35032.39	0.999	13577.85	5402.188	0.000229	1211.264
SVR	77,637.01	0.996	66,454.17	2,504.52	0.001	2,243.99
KNN	181,671.90	0.979	138,933.7	7,994.46	0.003	5,086.35
Decision tress	233,501.00	0.965	179,303.00	12,440.10	0.005	7,576.96

The years covered (2015–2020) are among the years Abu Dhabi ranked below its competitors in real estate transaction transparency. [4] Claims that rampant construction has a downward trend on house prices due to an oversupply of housing units are also supported; the decline in house prices over the period can also be attributed to an oversupply of housing units.

Similar to [7], this study implemented several predictive machine learning models to predict house properties in Abu Dhabi. The best model (Random Forest) had a higher R squared, indicating that it can account for a greater amount of variation in prices than the models built in [7].

5. Conclusion

From the analysis, we can conclude that original selling prices have the most impact on current house prices; this is attributable to the fact that original house prices are mostly set based on construction costs. The original selling price forms the basis of the original house value, which mutates to the current cost through depreciation or appreciation. Total units also impact the current house prices positively in Abu Dhabi; this can be attributed to the fact that big houses consume a large amount of land and building materials, which increase building costs.

In this research, decision trees, random forests, support vector regressions, and k-nearest neighbors were applied. The random forest model had the lowest RMSE and highest R-squared value of all the models tested.

6. Disclosure and conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] L. Igual and S. Seguí, “Introduction to Data Science,” in *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, L. Igual and S. Seguí, Eds. Cham: Springer International Publishing, 2017, pp. 1–4. doi: 10.1007/978-3-319-50017-1_1.

[2] P. C. Jackson, *Introduction to Artificial Intelligence: Third Edition*. Courier Dover Publications, 2019.

[3] E. Alpaydin, *Introduction to Machine Learning, fourth edition*. MIT Press, 2020.

[4] H.-A. Al-Malkawi and R. Pillai, “The impact of financial crisis on UAE real estate and construction sector: Analysis and implications,” *Humanomics*, vol. 29, May 2013, doi: 10.1108/08288661311319184.

[5] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, “Identifying Real Estate Opportunities Using Machine Learning,” *Appl. Sci.*, vol. 8, no. 11, Art. no. 11, Nov. 2018, doi: 10.3390/app8112321.

[6] A. S. Ravikumar, “Real Estate Price Prediction Using Machine Learning,” Masters thesis, Dublin, National College of Ireland, 2017. Accessed: Sep. 16, 2022. [Online]. Available: https://norma.ncirl.ie/3096/

[7] B. Park and J. K. Bae, “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2928–2934, Apr. 2015, doi: 10.1016/j.eswa.2014.11.040.

[8] Q. Truong, M. Nguyen, H. Dang, and B. Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” *Procedia Comput. Sci.*, vol. 174, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[9] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, “House Price Prediction using Random Forest Machine Learning Technique,” *Procedia Comput. Sci.*, vol. 199, pp. 806–813, 2022, doi: 10.1016/j.procs.2022.01.100.

[10] Imran & Zaman, Umar & Waqar, Muhammad & Zaman, Atif, "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data," *Fundamental Informaticae,* Vol. 1, pp. 11-23, 2021. 10.22995/scmi.2021.1.1.03.

[11] L. Breiman, Ed., *Classification and regression trees*, CRC Press repr. Boca Raton, Fla.: Chapman & Hall/CRC, 1998.

[12] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/BF00993309.

[13] T. M. Therneau, E. J. Atkinson, and M. Foundation, “An Introduction to Recursive Partitioning Using the RPART Routines,” *Mayo Found. Tech. Rep.*, vol. 61, p. 60, 1997.

[14] “Decision Tree Regression.” https://www.saedsayad.com/decision_tree_reg.htm (accessed Mar. 11, 2023).

[15] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[16] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[17] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[18] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[19] N. Cristianini, J. Shawe-Taylor, and D. of C. S. R. H. J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[20] P. Cunningham and S. J. Delany, “k-Nearest Neighbour Classifiers - A Tutorial,” *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.