Dubai Real Estate Price Prediction

Johnson
George Mason University
Fairfax, Virginia
kjohn30@gmu.edu

Myers
George Mason University
Fairfax, Virginia
tmyers24@gmu.edu

***Abstract* — This project focused on identifying key predictors of real estate property values in Dubai's dynamic market, offering crucial insights for investors, families, construction companies, and real estate agents. By examining rental and transactional data, the project addressed data quality issues and standardized units. The analysis determined that property area, number of beds, and property type are the most influential factors in predicting rental values. Results showed moderate performance of the model on rental data but limited effectiveness on transactional data, highlighting that property-specific features are more reliable indicators of value than location-related factors.**

Keywords— Random Forest Regression, R-Squared, Hyperparameters, Interquartile Range

I. Introduction

Dubai, the most populous city in the United Arab Emirates, is home to one of the world's most luxurious real estate markets. With tens of thousands of people moving to Dubai each year, it's important for stakeholders to understand which unique factors influence property price and popularity. These stakeholders can include investors, families, construction companies, and real estate agents interested in expanding in the area. While making an estimated guess is an option to examine the state of the Dubai real estate market, there is a much more reliable method. A computational model can offer great insights into not only a better understanding of Dubai's real estate, but also predicting future prices in the area. More specifically, a predictive model can be an extremely valuable tool in decision-making.

Recent studies show machine learning's effectiveness in housing price predictions. Nguyen and Xu (2022) found XGBoost best for Chicago, Baldominos et al. (2021) highlighted random forests and neural networks in Boulder, and Choy and Ho (2023) showed random forests outperforming hedonic models in Hong Kong. Overall, the random forest regression model proved to perform the best when predicting prices, validated by its performance measures. The purpose of this project is to develop and use a random forest regression model paired with hyperparameter tuning to determine factors that affect real estate prices in Dubai.

II. Methods

For the model, two historical datasets were used to test and explore it. The first is the "Dubai Real Estate, UAE Rental Market," which provides over 70,000 entries of rental property listings across various major cities in the United Arab Emirates, including Dubai. Compiled from Bayut, the UAE's largest real estate portal, the dataset includes 17 variables that describe the property's features and the annual rent price. The second dataset is the "Dubai Real Estate Transactions Dataset," with over 1,000,000 observations of real estate transactions in Dubai. Derived from Dubai Pulse, the data includes 48 variables with location-specific and price variables for deeper analysis. Using both datasets is important because while the first provides numerical data, the second offers useful categorical data to predict real estate prices.

Cleaning and preprocessing the raw datasets included removing outliers using the interquartile range (IQR) method, handling null and missing values, along with converting to U.S. standard unit types. **Fig. 1** displays the dataset's right-skewedness, meaning that the mean is greater than the mean. For the numerical Bayut dataset, exploratory data analysis determined that the 'number of beds', 'number of baths' (**Fig. 2**), 'property type' (**Fig. 3**), 'area in square feet' (**Fig. 4**), and 'listing days' variables were the most correlated with the annual rent price.

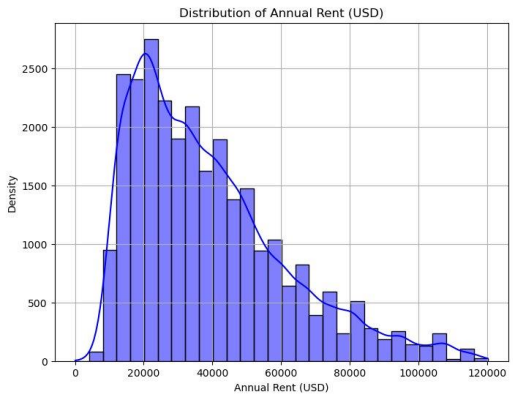

Fig 1: Histogram and Density Plot depicting the distribution of the annual rent variable in USD, displaying its right-skewedness

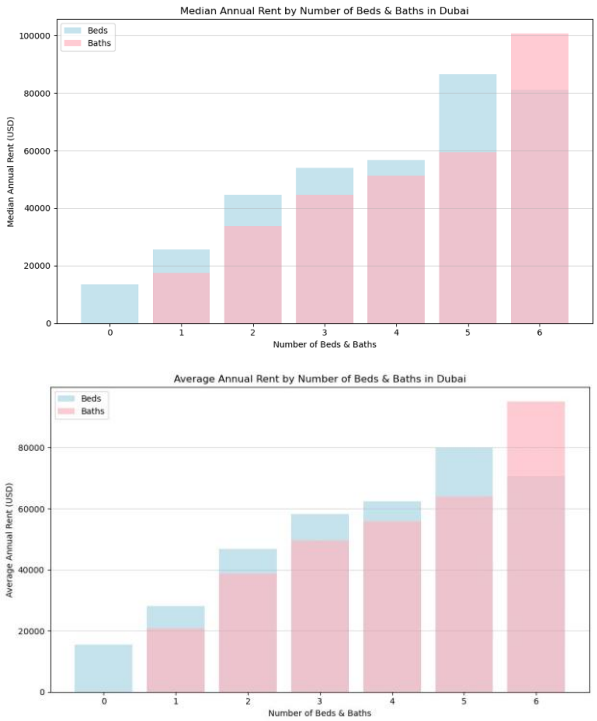

Fig 2: Two stacked bar charts displaying the average (top) and median (bottom) annual rent price in USD grouped by the count of beds and bathrooms. Both variables impact annual rent prices.

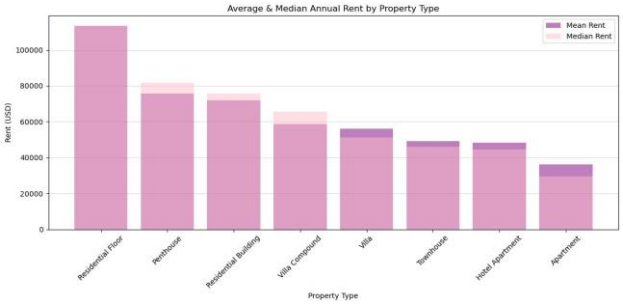

Fig 3: Bar chart displaying the average and median annual rent price in USD grouped by each individual property type. The 'Residential Floor' and 'Penthouse' types demand the highest prices, while the 'Townhouse' and 'Apartment' types demand the lowest. Property type influences rent price.

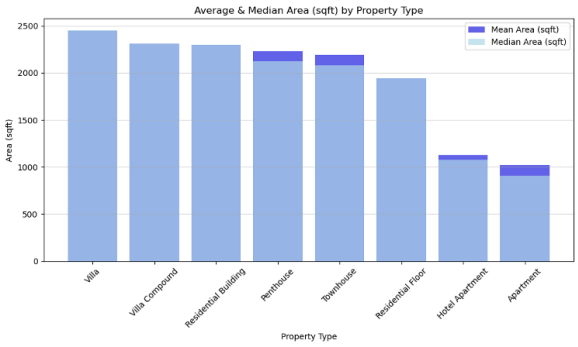

Fig 4: Stacked bar chart displaying the average and median property area in square feet by the property type. The 'Villa', 'Villa Compound', and 'Residential Building' types boast the highest square footage, while the 'Hotel Apartment' and 'Apartment' have the lowest. This confirms the relationship where the property type influences the area, which then impacts the annual rental price.

For the categorical Dubai Pulse dataset, the 'nearest landmark' (Fig 5), 'nearest metro station' (Fig 6, 'nearest mall' (Fig 7), 'rooms' (Fig 8), 'property type' (Fig 9), 'property sub-type' (Fig 10), and 'parking availability' (Fig 11) influenced the property price.

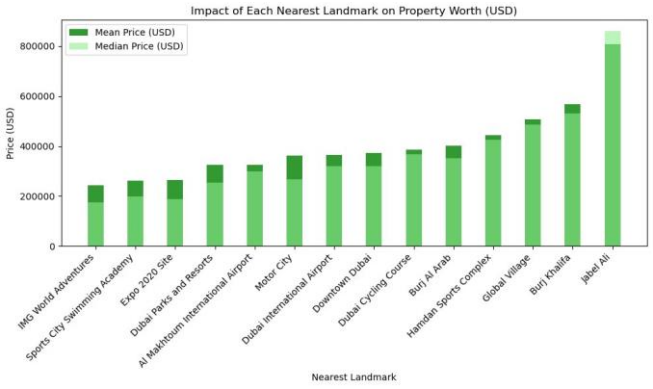

Fig 5: Stacked bar chart displaying the average and median price in USD grouped by the nearest landmark. Properties close to the 'Jabel Ali' and 'Burj Khalifa' landmarks demand the highest prices, while properties close to 'IMG World Adventures' and 'Sports City Swimming Academy' are lower.

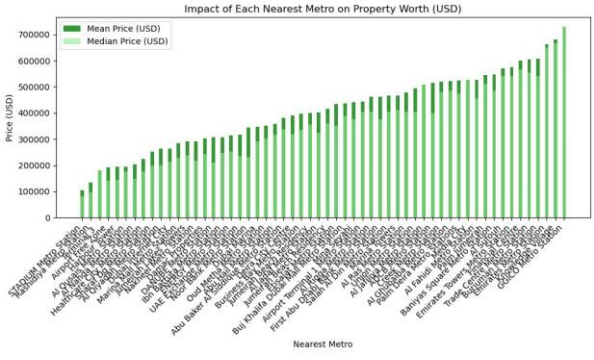

Fig 6: Stacked bar chart displaying the average and median price in USD grouped by the nearest metro station. The 'GGICO Metro Station', 'Knowledge Village', and the 'Emirates Metro Station' have the priciest properties nearby, while 'STADIUM Metro Station' and 'Rashidiya Metro Station' have the cheapest.

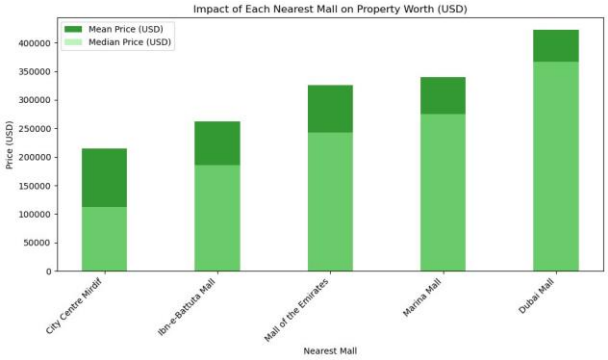

Fig 7: Bar chart displaying the average and median price in USD grouped by the nearest mall. Real estate close to 'Dubai Mall' and 'Marina Mall' are pricier, while 'City Centre Mirdif' and 'Ibn-e-Battuta Mall' are on the cheaper side.

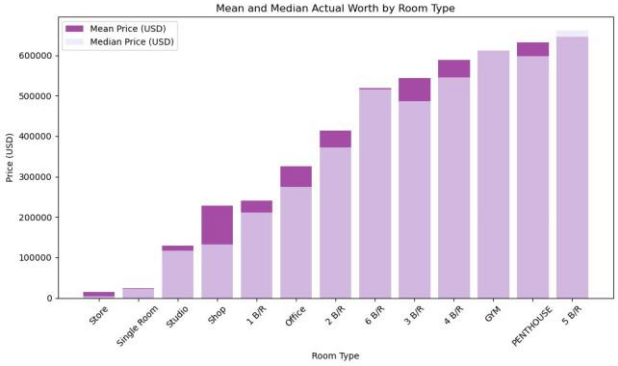

Fig 8: Bar chart displaying the average and median price in USD grouped by the room type and distribution. The most expensive is the '5B/R', 'PENTHOUSE', and 'GYM', while the cheapest is 'Store', 'Single Room', and 'Studio.'

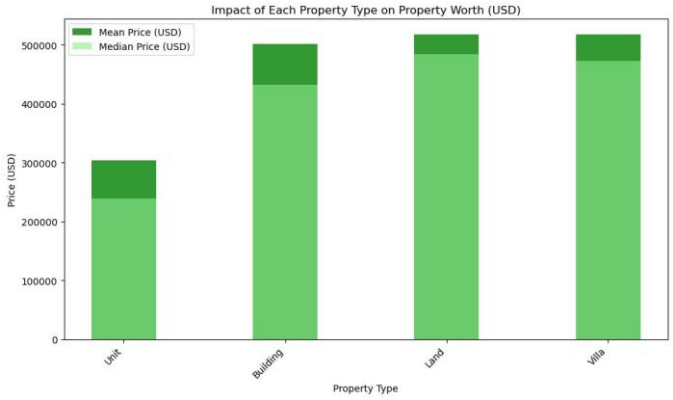

Fig 9: Bar chart displaying the average and median price in USD grouped by the property type. Both the 'Land' and 'Villa' types are the priciest, while the 'Unit' property type is the cheapest.

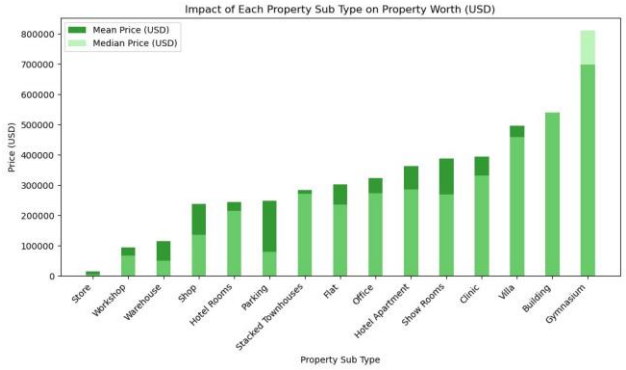

Fig 10: Bar chart displaying the average and median price in USD grouped by the property sub-type. 'Gymnasium' and 'Building' sub-types are most expensive, while the 'Store' and 'Workshop' subtypes are less.

Fig 11: Bar chart displaying the average and median price in USD grouped by whether the property has parking availability. Unexpectedly, properties with no parking are priced higher than those that do not have parking spaces.

Given this information, the models were constructed with a random forest regression algorithm. This approach involves creating multiple data subsets and then predicting each subset. These predictions are then averaged to obtain the final prediction. Once proper feature selection is conducted, the input parameters are incorporated into the random forest model for each dataset. Beforehand, cross-validation (CV) and grid search were used to implement hyper parameter tuning. This process involves specifying a range of hyperparameters and training the model with 20 percent of the historical data for each combination. The most efficient set of hyperparameters is selected, based on the CV performance. Once the model is trained, it is ready to predict future real estate prices. The model's performance is validated using Mean Squared Error (MSE), Root Mean-Squared Error (RMSE), and R-Squared values. The individual R-Squared values were assessed by measuring the impact of each feature's removal on the overall R-Squared score.

III. Results

The model using the numerical Bayut rental properties dataset produced mediocre results, yielding a training R-Squared of 0.67 and a test R-Squared of 0.623. This small difference indicates that the model is not underfit or overfit. Table I lists all the performance measures. Figure 12 shows the

features with the highest R-Squared values. Only two parameters proved to significantly contribute to the R-Squared of the model at values above 0.05. 'Area in square feet' was the largest contributor at 0.53, and the 'number of beds' was the second largest at 0.33. 'Listing age' was just below the threshold with a of 0.047. Figure 13 shows the relationship between the predicted and actual rent.

TABLE I. PROPERTIES DATASET PERFORMANCE MEASURES

	R-Squared	MSE	RMSE
Training	0.672	132,501,741	11,510
Test	0.623	152,679,217	12,35

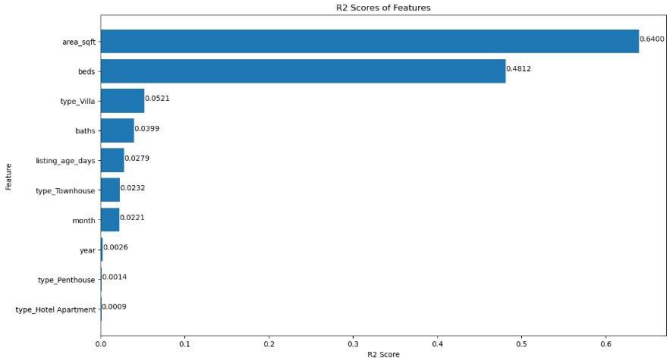

Fig 12: Horizontal bar chart that displays the R-Squared values for the model's input features. This shows that both area in square feet and number of beds contribute the most to the score. Property type also contributes, but comparatively less than the previous features.

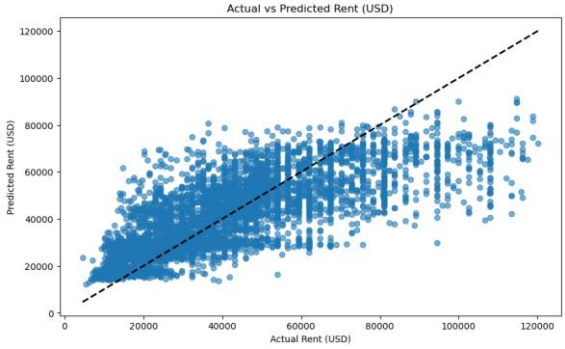

Fig 13: Scatterplot that shows the relationship between the predicted and actual annual rents. The scatterplot shows that the actual and predicted rents match reasonably well initially. However, the predictions become worse as the actual rent increases.

The model for the categorical Dubai Pulse dataset performed less than average, with a training R-Squared of 0.37 and test R-Squared of 0.35. These values indicate that the model is neither underfit nor overfit. Table II lists the performance measures. The 'property type', 'nearest metro station', and 'nearest landmark' predictors have an R-Squared value above 0.05. However, these R-Squared values are much lower than the best features in the previous model. Figure 14 shows the features with the highest R-Squared values. Figure 15 shows the relationship between the predicted and actual property values.

TABLE II. TRANSACTIONS DATASET PERFORMANCE MEASURES

	R-Squared	MSE	RMSE
Training	0.409	33,790,156,436	183,820
Test	0.408	33,941,817,845	184,233

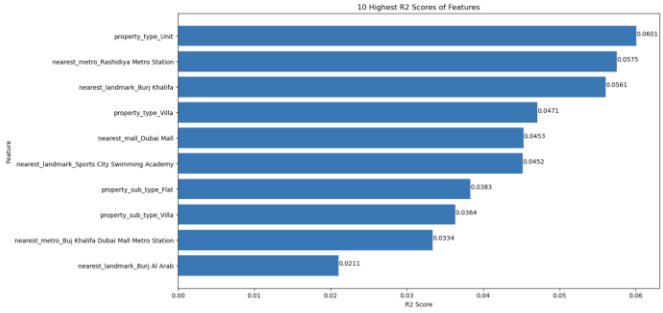

Fig 14: Horizontal bar chart that displays the 10 highest R-Squared values. This shows that property type contributes to the R-squared score the most. The nearest metro station and nearest landmark also contribute a significant amount to predicting value.

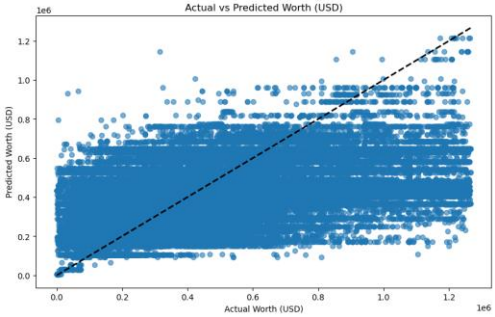

Fig 15: Scatterplot that shows the relationship between the predicted and actual annual property values. The scatterplot shows that the actual and predicted rents match poorly and that the model is poor at predicting property value.

IV. DISCUSSION

The results provide valuable insights into which property features are effective predictors of real estate prices in Dubai. As expected, a property's area is the most effective predictor of housing prices, along with the property type, given their own relationship with each other. However, the results revealed some

unexpected findings regarding other features of interest. Specifically, the ineffectiveness of many location-specific variables within the model was surprising. Contrary to common belief, the nearest metro station and nearest landmark were only marginally effective predictors of property value. In addition, Nearest mall did not significantly predict the real estate value. This is unexpected because of the widespread assumption that location and accessibility are two of the most important factors when determining the value of a property. These findings improve our understanding of which factors influence real estate prices in Dubai, along with those that do not. The results highlight that certain physical attributes of a property, such as area, the number of beds, and property type are critical predictors for real estate prices. At the same time, the results stress that location and proximity-related features are not as influential as traditionally believed. This insight could potentially prompt a reevaluation of the emphasis placed on location within Dubai real estate and suggest a shift to more impactful predictors. This understanding can inform future models and strategies within the Dubai real estate industry.

V. REFERENCES

[1] Nguyen, H., & Xu, K. (2022). Predicting housing prices and analyzing real estate market in the Chicago suburbs using machine learning. arXiv preprint arXiv:2210.06261. Retrieved from https://arxiv.org/abs/2210.06261

[2] Baldominos, A., Blanco, I., Moreno, A. J., & Afonso, C. (2021). Machine learning, deep learning, and hedonic methods for real estate price prediction. arXiv preprint arXiv:2110.07151. Retrieved from https://arxiv.org/pdf/2110.07151

[3] Choy, L. H. T., & Ho, W. K. O. (2023). The use of machine learning in real estate research. Land, 12(4), 740. https://doi.org/10.3390/land12040740