Using ML Algorithms to Predict Real Estate Units Valuations in Ajman

Heba Al Yazouri
haa5235@rit.edu

Follow this and additional works at: https://repository.rit.edu/theses

Recommended Citation

This Master’s Project is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.
Using ML Algorithms to Predict Real Estate Units Valuations in Ajman

by

Heba AL Yazouri

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
January 2023
Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: **Heba AL Yazouri**
Graduate Capstone Title: **Using ML Algorithms to Predict Real Estate Units Valuations in Ajman**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak
    Date:

    Chair of committee

Name: Dr. Boutheina Tlili
    Date:

    Member of committee
Acknowledgments

Words cannot express my gratitude to my mentor Dr. Boutheina Tlili and chair of committee Dr. Sanjay Modak, for their generous support and invaluable patience and feedback on my project.

I am also grateful to Mr. Omar bin Omar the Director-General of the Department of Land and Real Estate Regulation, and my direct manager Ms. Amna Al Ali the Head of the IT Section, who’s I also could not have undertaken this project without their support and allowing me using the data. Thanks, should also go to the Engineering Section, especially Mr. Ayman Hassan, for his support with data collection.

Lastly, I extend my heartfelt thanks to my family, especially my parents. Their support and belief in me kept my motivation high during this journey.
Abstract

This project applies machine learning algorithms to predict the real estate unit’s values in the emirates of Ajman. Adopting machine learning techniques to predict real estate unit prices helps us produce more accurate valuations. It allows us to use historical data that can’t be used in traditional valuation techniques.

Datasets from the Department of Land & Real Estate were used to run an exploratory analysis to explore the most critical factors influencing real estate unit prices in Ajman. It shows that the unit area is the most significant factor affecting Ajman's real estate unit prices. It also shows how the real estate market is trending and the effect of Covid-19 on the real estate market in Ajman.

Predictive Analysis was applied using the R programming language to predict real estate unit prices. Three models were used: Multiple Linear Regression, Support Vector Regression (SVR), and Gradient Boosting (GB) Model. Stepwise Regression was used to select the most significant attributes that will be used to build the model. Genetic Algorithms were used to tune parameters for the Support Vector Regression and Gradient Boosting Model. The resulting models were then evaluated using R2 and RMSE, where the GB model provided the best results.

This project will help Ajman Land and Real Estate Regulation Department to get more accurate values for real estate unit prices on the spot, which reduces the service time and steps. It also helps create a more reliable real estate market.

Keywords: Ajman, Real Estate Valuation, Machine Learning, Support Vector Regression, Multiple Linear Regression, and Gradient Boosting (GB) Model
## Table of Contents

ACKNOWLEDGMENTS .................................................................................................................. II

ABSTRACT ................................................................................................................................... III

LIST OF FIGURES ........................................................................................................................ VI

LIST OF TABLES ........................................................................................................................ VII

CHAPTER 1 - FOUNDATION .................................................................................................1

1.1 INTRODUCTION ............................................................................................................1

1.2 PROJECT STATEMENT .................................................................................................1

1.3 PROJECT GOALS ...........................................................................................................2

1.4 AIMS AND OBJECTIVES ..............................................................................................2

1.5 RESEARCH METHODOLOGY .......................................................................................2

1.6 LIMITATIONS OF THE STUDY ....................................................................................3

CHAPTER 2 – LITERATURE REVIEW ....................................................................................4

CHAPTER 3- PROJECT DESCRIPTION ..................................................................................9

3.1 DATA ANALYTICS TOOLS AND SOFTWARE USED ......................................................9

3.2 DATA ANALYSIS MODELS USED ................................................................................9

3.2.1 MULTIPLE LINEAR REGRESSION .........................................................................9

3.2.2 SUPPORT VECTOR REGRESSION ........................................................................10

3.2.3 GRADIENT BOOSTING MODEL ..........................................................................11

CHAPTER 4- DATA ANALYSIS ...........................................................................................13

4.1 BUSINESS UNDERSTANDING ....................................................................................13

4.2 DATA UNDERSTANDING .............................................................................................13

4.2.1 DATA COLLECTION ....................................................................................................13

4.2.2 DATA DESCRIPTION ................................................................................................14

4.2.3 DATA EXPLORATION ...............................................................................................16

4.3 DATA PREPARATION .....................................................................................................20

4.3.1 DATA PREPROCESSING ........................................................................................20

4.3.2 DATA EXPLORATION AND VISUALIZATION .........................................................25

4.3.2.1 WHAT ARE THE HISTORICAL TRENDS FOR UNIT PRICES IN AJMAN? AND WHAT IS THE EFFECT OF COVID-19 ON THE REAL ESTATE UNITS MARKET IN AJMAN? ..............25

4.3.2.2 WHAT ARE THE FACTORS AFFECTING REAL ESTATE UNITS’ PRICES? ..........27

4.4 MODELING ....................................................................................................................32
4.4.1  STEPWISE REGRESSION FOR FEATURE SELECTION ........................................................ 33
4.4.2  MULTIPLE LINEAR REGRESSION .................................................................................. 34
4.4.3  SUPPORT VECTOR REGRESSION .................................................................................. 36
4.4.4  GRADIENT BOOSTING MODEL ................................................................................... 37
4.5  EVALUATION .................................................................................................................... 39
4.6  DEPLOYMENT .................................................................................................................... 41
CHAPTER 5 - CONCLUSION .................................................................................................. 42
5.1  CONCLUSION ................................................................................................................... 42
5.2  RECOMMENDATIONS AND FUTURE WORK ................................................................. 42
REFERENCES ......................................................................................................................... 43
List of Figures

Figure 1: Illustrative Example of the SVR model (Sharp, 2020) ................................................................. 11
Figure 2: Gradient Boosting Algorithm Equations (Masui, 2022) .............................................................. 12
Figure 3: 'Units Sales' Dataset .................................................................................................................. 14
Figure 4: 'Units View' Dataset .................................................................................................................. 14
Figure 5: 'Projects Completion Dates' Dataset .......................................................................................... 14
Figure 6: Number of rows and columns for the Units Sales Dataset .......................................................... 16
Figure 7: Data types of Units Sales Dataset attributes ................................................................................. 16
Figure 8: Numerical attributes summary statistics for the Units Sale dataset ........................................... 17
Figure 9: Categorical attribute summary statistics for the Units Sales dataset ........................................... 17
Figure 10: Sale Date attribute summary statistics ...................................................................................... 18
Figure 11: Number of rows and columns in the ‘Units View’ dataset ......................................................... 18
Figure 12: Data Types of Units View dataset attributes .............................................................................. 18
Figure 13: Summary statistics about the Units View dataset attributes ...................................................... 19
Figure 14: Number of rows and columns in the Projects Completion Dates dataset .................................. 19
Figure 15: Data Types of Projects Completion Dates dataset attributes .................................................... 19
Figure 16: Summary statistics about the Projects Completion Dates dataset ............................................ 20
Figure 17: Distribution of rooms count over Units Areas for each Unit Type .............................................. 22
Figure 18: boxplots of the sale price and unit area before removing outliers ............................................. 22
Figure 19: boxplots of the sale price and unit area after removing outliers ................................................. 23
Figure 20: Summary statistics for the ‘view’ attribute in the joined dataset .............................................. 23
Figure 21: Summary statistics for the ‘completion date’ attribute in the joined dataset ............................. 24
Figure 22: Summary statistics for the ‘unit age’ attribute .......................................................................... 24
Figure 23: Average unit prices from 2017 Q1 to 2021 Q1 ......................................................................... 26
Figure 24: Monthly sum of unit prices for the year 2017 to 2020 .............................................................. 26
Figure 25: Average unit prices per district and room count ...................................................................... 27
Figure 26: Distribution of average unit prices over the unit area (in sqm) per room count ............................. 28
Figure 27: Average unit price and unit age per district and room count ..................................................... 28
Figure 28: Average unit price per unit view ............................................................................................... 29
Figure 29: Average unit price per district and unit view .......................................................................... 29
Figure 30: Average unit price and unit age per district and unit view ......................................................... 30
Figure 31: Average unit prices and unit areas for each floor number ......................................................... 31
Figure 32: Average unit price per rooms count and bathrooms count ...................................................... 31
Figure 33: Correlation plot for the numerical attributes ............................................................................ 32
Figure 34: The output model for Stepwise Regression ................................................................................ 34
Figure 35: The output model for Multiple Linear Regression ..................................................................... 35
Figure 36: MLR model performance measures results ............................................................................ 35
Figure 37: Genetic Algorithm result for SVR parameters ........................................................................ 36
Figure 38: The output model for the SVR ................................................................................................. 36
Figure 39: SVR model performance measures results ................................................................................ 37
Figure 40: Genetic Algorithm result for GB model parameters ................................................................. 38
Figure 41: The output model for the GB ..................................................................................................... 38
Figure 42: Attributes influencing the GB model .......................................................................................... 38
Figure 43: GB model performance measures results .................................................................................. 39
Figure 44: Actual vs. predicted sale prices using the MLR model ............................................................... 40
Figure 45: Actual vs. predicted sale prices using the SVR model ............................................................... 40
Figure 46: Actual vs. predicted sale prices using the Gradient Boosting Model ........................................... 41
List of Tables

Table 1: 'Units Sales' Dataset Description .................................................................................................................... 14
Table 2: 'Units View' Dataset Description .................................................................................................................. 15
Table 3: 'Projects Completion Dates' Dataset Description ........................................................................................ 16
Table 4: Models performance measures results ..........................................................................................................39
Chapter 1 - Foundation

1.1 Introduction

The Department of Lands and Real Estates Regulation in Ajman is responsible for registering, organizing, and promoting the real estate market in the emirates Ajman by setting the policies, executing the projects and initiatives, and applying the best international practices in the field. The department aims to provide an attractive and reliable real estate investment and development environment by issuing new laws and legislation, providing the customers and investors with the appropriate facilities and services, in addition to monitoring and controlling the real estate market prices so they are validating the price of each sale transactions to ensure that it is aligning with the real estate market values to protect investors from being fooled by underestimating their properties prices, furthermore to prevent the loss of the government fees as the sale fees is a percent of the property price.

The real estate market in Ajman can be categorized into:

- Lands: that includes the non-built lands
- Real estate properties: that includes the built lands like (villas, buildings, etc.)
- Real estate units: these are units within the real estate development projects like apartments, studios, shops, and villas. In this project, we are focused on apartments and studios.

The employment of data analytics in this sector greatly helps determine a data-driven market price of real estate properties. So, in this project, we studied the factors that affect the real estate units’ market prices using data analytics techniques. In addition, we used machine learning algorithms to predict the real estate market prices for the real estate units using the historical transactional data available in the Lands and Real Estate Department to help in providing a reliable real estate investment environment.

1.2 Project Statement

Manually evaluating the real estate unit in Ajman Land and Real Estate Regulation Department consumes a long time that increases the services time; in addition, it's not taking into consideration the historical data. So, in this project, we used data-driven models to predict the real estate unit's market value based on the previous sales prices that will reduce the service time and steps and provide accurate unit valuations.
Additionally, in this project, we applied several data analytics techniques and algorithms to analyze the historical sales prices and answer the following research questions:

- What are the historical trends for unit prices in Ajman? And what is the effect of Covid-19 on the real estate units’ market in Ajman?
- What are the factors affecting real estate units’ prices?
- How accurate is the market price predicted using data mining algorithms?

1.3 Project goals

In this project, we used several machine learning algorithms to analyze the historical sale price data and predict the unit prices. This aims to produce more accurate real estate unit valuations on the spot, reducing the time and steps for Ajman Land and Real Estate Regulation Department services and providing a more reliable real estate market.

The datasets used had information about the real estate units’ details and geographic information such as unit area, number of bedrooms, unit view, etc... Data analytics techniques were used to understand the relationship between the unit price and other independent attributes to identify the factors that affect real estate units’ valuation of the Real estate development projects in the emirates of Ajman.

1.4 Aims and Objectives

- Perform exploratory analysis to identify the relationship between the sale prices and other independent attributes.
- Implement a Stepwise Regression model to identify the most significant attributes in predicting the real estate unit prices.
- Implement three machine learning models, which are Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Gradient Boosting (GB), to predict the real estate unit prices.
- Evaluate the built models and identify the most accurate model in predicting the real estate unit prices.

1.5 Research Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was implemented in this project to achieve the project goals and objectives. The below CRISP-DM phases were followed:
• Business Understanding: We started this project by understanding the business domain and identifying the project goals and objectives. The project aims to implement machine learning models to predict a data-driven market price of real estate units that could reduce the service time and produce more accurate valuations. In addition, data analytics techniques were used to understand the most significant factors affecting the real estate units in Ajman.

• Data Understanding: The needed datasets were collected from the Department of Lands and Real Estate Regulations. Then the data were described and explored to understand the data and examine the quality of the data.

• Data Preparation: Multiple data preprocessing tasks were performed to prepare the final dataset that can be used in the subsequent modeling phase. Data preprocessing tasks include data cleaning, feature engineering, feature selection, and data transformation.

• Modeling: Three machine learning models were implemented, which are Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Gradient Boosting (GB) using the training dataset. The models were tested and assessed using the testing dataset.

• Evaluation: The resulting models were evaluated to choose the best model in terms of accuracy and ability to fit the testing dataset.

• Deployment: The chosen model will be deployed to the staging environment, tested with recent data, and then deployed to the live environment.

1.6 Limitations of the Study

The following limitations were identified through this project that can be considered in future work:

• The project was considering the residential real estate units and limited to the apartment and studio unit types. Villas and commercial units were not considered due to the limited data about them.

• The analysis was based only on units’ data and historical sale prices in the Department of Lands and Real Estate Regulations. It didn’t include other factors affecting real estate valuations, such as nearby establishments and services in the location and population data.
Chapter 2 – Literature Review

The real estate sector is one of the oldest emerged sectors in human civilization. Moreover, it is a crucial sector for the growth of the countries’ economy as real estate market values are an indicator of the country's economic status, welfare, and stability. (Al-Hamadin & Al-Sit, 2020)

Ajman is one of the smallest emirates in the UAE but is developing readily. The real estate sector represented 12.5% of the GDP (gross domestic product) of the Emirate of Ajman in 2020 and 11.9% in 2021. (Ajman Statistics and Competitiveness center, 2021). Ajman was one of the first emirates that issued the Real estate freehold Law in UAE, which allows foreigners to buy land or property in Ajman and register it with his ownership, so he has the absolute right to sell, lease or live on the property. The introduction of this law resulted in instant hikes in property prices (Ajman Real estate Freehold Law, 2021). It also issued several regulations to regulate real estate development projects to provide a reliable and attractive real estate investment environment (Marlow, 2012).

(Mishra, 2021) highlights that there is a difference between the market price of a property and the market value of a property. The market price is the price that the buyer paid towards buying the property; in other words, it is the actual price of the sale transaction, while the market value is how much the property is worth in the market, mainly from the buyer's perspective.

According to (Folger, 2021), real estate valuation is needful for various reasons, including financing, investment analysis, and property insurance. Three Different approaches are used to appraise a real estate value: the sales comparison approach, cost approach, and income capitalization approach. The sales comparison approach is most suitable for evaluating the value of residential lands and real estate units. The estimation is based on comparing the recent sale prices of properties with similar characteristics.

The property market value is affected by critical factors such as the property location, quality of building construction, the age of the property, and the project developer brand, as the well-known developers' project prices will be more than starters developers' project prices (Mishra, 2021). Other characteristics to consider in property valuation is the condition of buildings, date of sale, and sales price, in addition to some physical features such as plot area, landscaping (Folger, 2021), number of bedrooms, number of bathrooms, number of rooms, utility available, and property view. (Khare, Gourisaria, Harshvardhan, Joardar, & Singh, 2021).
(Pai & Wang, 2020) examined the effect of feature selection on the machine learning algorithms results by building four different machine learning models to predict the sale prices two times: one time using all available attributes and the other with selected attributes. The results showed that the models with the eleven selected significant attributes produced more accurate results. Pearson correlation coefficient was used to determine the significant features; they decided that significant attributes are with an absolute correlation coefficient larger than 0.1. On the other hand, (Khare, Gourisaria, Harshvardhan, Joardar, & Singh, 2021) examined also building different machine learning algorithms one time using 18 predictor features and the other using only two predictor features which are the highly correlated features with the property price. The study result was that the predictions of the first model’s 18 features were better than the second model. We can conclude that property valuation using machine learning algorithms depends on various features combined and not just a few correlated features, so feature selection is a crucial phase and significantly affects the prediction results of the machine learning models.

The traditional valuation methods are based on the current state of the property and real estate market and do not allow the appraiser to use historical information. Data mining algorithms will enable us to use the power of historical market prices to identify the changes in the price level in a specific area, which will result in more precise valuations. (Hromada, 2016).

Adopting machine learning techniques and predictive analytics to evaluate real estate values will help us produce more accurate valuations. (Al-Hamadin & Al-Sit, 2020) Uses in their study several predictive models, which are: Linear Regression, GB-Regression, SVM, Random Forest, and MLP, to predict the real estate property values in Amman. According to their evaluation, MLP, Linear Regression, and GB-Regression generated the best results. At the same time, other algorithms created over-fitted models of the training dataset, which can be due to the use of small datasets that may have a low number of comparable instances.

According to the study conducted by (Trawinski, et al., 2017) that comparing the accuracy of expert algorithms with machine learning models for real estate appraisal, machine learning models were more accurate in predicting real estate values. The expert algorithms were based on a sales value comparison approach with similar properties. They have used Three different ML algorithms, Pruned Model Tree (M5P); which employs decision trees, Multilayer Perceptron; which is a feed-forward neural network employing the backpropagation learning algorithm, and Linear Regression; which is a statistical approach to create a linear function that represents the relationship between the predictor variables and the output variable. Pruned
Model Tree produced the best results among other ML algorithms. The predictors' variables used were the area, year of building construction, number of floors in the building, and number of rooms in the flat, including kitchens.

(Masías, et al., 2016) Conducted a comparison between several ML algorithms, which are Random Forest, Support Vector Machine, Neural Network, and Multiple Linear Regression, in terms of their performance in predicting Santiago residential property prices. They have segmented the dataset to 70% as a training dataset and 30% as a testing dataset and used these datasets for all ML algorithms to ensure a fair comparison. In addition, root-mean-squared-error (RMSE), mean-absolute-error (MAE), and the Pearson coefficient of determination (R2) were used as performance measurements for the selected ML algorithms. Based on this experiment, they concluded that the RF algorithm produced the best outcomes in predicting Santiago residential property prices, followed by SVM, MLR, and NN, respectively.

(Krishna, 2021) and (AlHathboor, 2020) developed property price prediction systems for the emirate of Dubai using machine learning algorithms. Both use different machine learning algorithms; however, Gradient Boost Regression produced the best results out of the rest models in both studies. (Krishna, 2021) built four prediction models: Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boost Regression. He used datasets from Dubai Land Department, RERA, and the Property Finder website. While (AlHathboor, 2020) built Decision Tree, Artificial Neural Network, and Gradient Boost Regression models. The dataset used contains the properties characteristics and prices data obtained from the Property Finder website and demographic data obtained from Dubai Statistics Center (DSC) and Google Maps database. GB model achieved an accuracy of 90.6%. Categorical attributes were encoded using label encoding based on the average price of each value in the categorical attribute, where the value with the lowest average price was assigned to 0, and the value with the highest average price was assigned to the highest number.

(Baldominos, et al., 2018) Used K-Nearest Neighbors, Support Vector Regression, Ensembles of Regression Trees, and Multi-Layer Perceptron algorithms to predict the property prices in the Salamanca district in Madrid (Spain). The study reported that Ensembles of Regression Trees outperform the other models, then K-Nearest Neighbors, Support Vector Regression, and the Multi-Layer Perceptron, respectively. Categorical features were transformed using One-hot encoding, where they create a binary feature for each value in the categorical features. This will allow the machine learning model to handle the categorical features. Five-fold cross-validation was used to avoid biased results when sampling the training and testing dataset. Cross-validation was also used by (Krishna, 2021) to avoid overfitting or underfitting the training
dataset to the model. He split the dataset into training and testing datasets and then applied k-fold cross-validation to the training dataset. First, the training dataset was divided into k datasets; then, these datasets were trained in k iterations; in each iteration, one of the split datasets was considered the testing dataset, and the rest were the training dataset. The average performance measure of the k folds is reported as a performance measure of the final model.

Each machine learning algorithm is configured based on a set of hyperparameters, and the suitable hyperparameters vary from one dataset to another. Different techniques can be used to tune the hyperparameters; (Baldominos, et al., 2018) manually tried a set of hyperparameters for each model, which was the default values or values commonly found in the literature. While (Pai & Wang, 2020) used a Genetic Algorithm to tune the model’s hyperparameters. The Genetic Algorithm is an optimization algorithm used to select the best-of-fit test solution by the rare and random mutation occurrence. The algorithm runs several models called populations with different sets of configurations called individuals or chromosomes that are initially selected randomly and then changed by crossover and mutation probability. This process is performed several times until we get the optimal solution of the objective function, each called a generation or iteration (Viadinugroho, 2021). In the study of (Pai & Wang, 2020), ten generations were performed with 20 populations. Crossover and mutation rates were 0.6 and 0.6, respectively, and average absolute percentage error was employed as the objective function to select the optimal hyperparameters for four machine-learning models-namely Least Squares Support Vector Regression (LSSVR), Classification and Regression Tree (CART), General Regression Neural Networks (GRNN), and Backpropagation Neural Networks (BPNN).

(Ravikumar, 2017) used different tools to develop several machine learning algorithms predicting real estate property prices. Random Forest, Multiple Regression, Support Vector Machine, Gradient Boosted Trees, Neural Networks, and Bagging algorithms were used and then evaluated using the following performance measurements:

- RMSE is the root-mean-squared error, which calculates the error rate in predicting the dependent variable and has the same unit as the dependent variable.
- R2: is the Pearson coefficient of determination, which is a percentage that calculates how well the model can explain the dependent variable's variance.

The best model was Random Forest and Gradient Boosted Trees, as they produced more accurate and less errors results.

This literature review guided me through my project and helped me successfully implement it, the most important outcomes were as follows:
- It helps me understand the project domain and the importance of real estate valuation.
- Through the literature review, I went through different studies that use different ML techniques to predict real estate property prices in various areas of the world. Each of them produced different results and performances for each ML algorithm. This gave me an overview of the most used ML algorithms that had the best results and helped me on deciding on the algorithms that were used in my project which are Gradient Boosting, SVM, and Multiple Linear Regression models.
- It introduced me to different features that affect real estate unit prices where each study was using different attributes. This in turn guided me in the data collection for my project where I added two features to be collected, which are the unit age and unit view.
- It showed me the importance of feature selection on the machine learning algorithms results, where I decided to use Stepwise Regression for feature selection in my project.
- Different techniques were used in the literature review for encoding the categorical attributes, and I decided to use the one-hot encoding technique rather than label encoding as there was no relation between the categorical attribute values in my dataset.
- Different techniques were used in the literature review for tuning the machine learning models hyperparameters, and I decided to use the Genetic Algorithm because I found it a more powerful technique.
Chapter 3- Project Description

In this project, we used three data sets collected from the department of lands & real estate regulation to analyze the factors affecting real estate unit market prices and predict real estate unit market prices using machine learning algorithms. The Cross-Industry Standards Process for Data Mining (CRISP-DM) methodology was implemented in this project. First, we initiate the project by understanding the real estate valuation domain and identifying the project’s business objectives. The next step was to understand and explore the collected data. Then, the final dataset was prepared by cleaning and joining the collected datasets to be ready for analysis and modeling. In the last step, the data were used to draw insights and build three machine learning algorithms: Multiple Linear Regression, Support Vector Regression, and Gradient Boosting Model.

3.1 Data Analytics Tools and Software Used

R was used in this project as an open-source programming language widely used in data analytics projects because it provides several libraries useful for analytics. RStudio was used as an integrated development environment (IDE) for R. It was used for data exploration and preprocessing, in addition to building and evaluating machine learning models.

Tableau software was used as an interactive data visualization software. It allows us to create powerful and valuable visualizations used to uncover some insights.

3.2 Data Analysis Models Used

Three predictive data analysis models were used in this project to predict the real estate unit prices.

3.2.1 Multiple Linear Regression

Multiple linear regression (MLR) is a regression model used to model the linear relationship between a continuous dependent variable and a set of independent variables that can be categorical or continuous, where we can find how the dependent variable is influenced by the independent variables simultaneously. The following equation can mathematically represent the MLR model (Simplilearn, 2022):
\[ Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \]

- \( Y_i \): is the dependent variable to be predicted.
- \( \beta_0 \): is the y-intercept. Which represents the value of \( Y_i \) when all other parameters are 0.
- \( \beta_1 x_1 \): is the value of the first independent variable. \( \beta_1 \) is the regression coefficient of the first independent variable, which represents the amount of change in \( Y_i \) when \( x_1 \) changes by one unit.
- \( \beta_k x_k \): is the regression coefficient by the value of the last independent variable.
- \( \varepsilon \): is the model’s error which is called residuals.

### 3.2.2 Support Vector Regression

Support Vector Regression (SVR) is an extension of classical support vector machine classifiers that can be used for regression problems. SVR was used because it allows us to model the linear or non-linear relationships between variables, giving the model more flexibility to fit the data. The SVR function can be a line, plane, hyperplane, or curve (for non-linear models) used to predict the continuous dependent variable (Kumar, 2018). Assuming SVR objective function is a line (\( y_i = w_i x_i \)), then the aim of the SVR model is to reduce errors by minimizing the predicted points outside the margins of the objective function and maximizing the predicted points inside the margins of the objective function (Dobilas, 2020) (as illustrated in Figure1). The SVR model should satisfy the following Equation (Sharp, 2020):

Minimize:

\[ \text{MIN} \quad \frac{1}{2} |w|^2 + C \sum_{i=1}^{n} |\xi_i| \]

Constraints:

\[ |y_i - w_i x_i| \leq \varepsilon + |\xi_i| \]
• $\varepsilon$: is the width of the margin or decision boundary lines around the SVR line or hyperplane function, called epsilon.
• $\xi$: is the amount of deviation of the predicted point from the margin of the SVR line or hyperplane function, called slack.
• C: is the penalty factors or cost of points predicted outside the margin of the SVR line or hyperplane function.

In addition, the following hyperparameters of the SVR model are essential to be tuned:

• The kernel function, which is the function that converts the data to higher dimensional data. (Dobilas, 2020)
• Curviness rate that we want for the decision boundary, which is called gamma (Kumar, 2018)

### 3.2.3 Gradient Boosting Model

Gradient Boosting (GB) is a Decision-tree-based algorithm that can be used in regression problems. Gradient Boosting uses boosting in building the tree, which means that trees are built one after the other, and each new tree should improve the insufficiency of the previous tree. Therefore, it will improve the accuracy of the overall model. The model starts with a constant value prediction for the dependent variable for example the mean, this prediction will produce results with high errors or residuals. The next step will be that the model will start building the first regression tree by finding the patterns between residuals of the first prediction and the independent features and adding to it a learning rate which is a constant number between 0 and 1 preventing overfitting the model to the training dataset. The new tree will reduce the errors by producing predictions with lower residuals. After that, the model will continue creating an improved regression tree one after the other.
utilizing the residuals of the previous iteration tree until the model accuracy stops improving. GB model can be illustrated as Equations in Figure 2. (Masui, 2022)

The following hyperparameters of the GB model are essential to be tuned (CFI Team, 2023):

- Loss function, which is the function used to decide the accuracy of the tree in each iteration.
- Number of trees or the number of iterations performed to build the final tree.
- Maximum depth of each tree, or the number of splits in each tree. This parameter controls the complexity of the model.
- Learning rate or shrinkage, which specifies the step size in each iteration or controls how fast the algorithm will be scaled down in each iteration that prevents overfitting the model to the training dataset.

**Gradient Boosting Algorithm**

1. Initialize model with a constant value:
   \[
   F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)
   \]

2. for \( m = 1 \) to \( M \):
   2-1. Compute residuals \( r_{im} = -\left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \) for \( i = 1, \ldots, n \)
   2-2. Train regression tree with features \( x \) against \( r \) and create terminal node reasons \( R_{jm} \) for \( j = 1, \ldots, J_m \)
   2-3. Compute \( \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \) for \( j = 1, \ldots, J_m \)
   2-4. Update the model:
   \[
   F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm})
   \]

Figure 2: Gradient Boosting Algorithm Equations (Masui, 2022)
Chapter 4- Data Analysis

This chapter provides an analysis of the collected datasets and then uses machine learning algorithms to predict the market price of a real estate unit. R programming language was used for data preparation and building the machine learning models, and Tableau was used to visualize and draw insights from the data.

4.1 Business Understanding.

The real estate unit is manually evaluated in the Department of Lands and Real estate Regulation by the valuation section either to validate the price of unit sale transactions or for personal valuation requests. The sales comparison approach is used to evaluate the residential real estate units’ value, where the estimation is based on comparing the recent sale prices of properties with similar characteristics.

The sale registration and real estate appraisal services are two of the most requested services in the Department of Land and real estate regulation. As a result, the number of transactions increased over the years; where the number of transactions was 5295 in 2020 and grew to 7218 in 2021. In addition, Ajman's total trading values for lands and real estate units reached 7,754,236,947 AED in 2021 compared to 4,581,835,616 AED in 2020 (Ajman Statistics and Competitiveness center, 2021).

In this project, we implemented machine learning models to predict a data-driven market price of real estate units that could reduce the service time and produce more accurate valuations. In addition, data analytics techniques were used to understand the most significant factors affecting the real estate units in Ajman.

4.2 Data Understanding

After understanding the domain of this project and setting our goals and objectives, we started exploring the datasets and understanding their attributes.

4.2.1 Data Collection

Three datasets were used in this project and collected from the Department of Land & Real Estate Regulations—Ajman Government. The first dataset, the ‘Units Sales’ dataset, contains historical real estate units’ sales transitions from January 2017 to April 2021, in addition to real estate units’ details and geographic information.
The ‘Projects Completion Dates’ second dataset contains information about the real estate projects’ completion dates. The last dataset was the ‘Units View’ dataset includes information on the view of the real estate units in each project, whether it is a garden view, street view, etc..., and this dataset was not ready, but we created it with the help of the engineers in the Lands Department. These datasets were joined and cleaned to produce the final dataset, as illustrated in the data preprocessing section.

4.2.2 Data Description

The below tables contain an overview and description of all the collected dataset’s attributes.

Table 1: ‘Units Sales’ Dataset Description

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Attribute (variable)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units Sales</td>
<td>Project Id</td>
<td>Id of the project that the unit listed belongs to.</td>
</tr>
<tr>
<td></td>
<td>Project Name</td>
<td>Name of the project that the unit listed belongs to.</td>
</tr>
<tr>
<td>Attribute</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Unit Number</td>
<td>The number of the listed unit</td>
<td></td>
</tr>
<tr>
<td>Unit Type</td>
<td>Type of the unit listed (apartment, villa, studio...)</td>
<td></td>
</tr>
<tr>
<td>Unit Type Id</td>
<td>Id of the type of the unit listed.</td>
<td></td>
</tr>
<tr>
<td>Sale Price</td>
<td>The sale transaction price of the listed unit.</td>
<td></td>
</tr>
<tr>
<td>Sale Date</td>
<td>Sale transaction date of the listed unit (from Jan 2017 to Apr 2021).</td>
<td></td>
</tr>
<tr>
<td>Unit Use</td>
<td>Use of the unit listed (residential or commercial).</td>
<td></td>
</tr>
<tr>
<td>Floor Number</td>
<td>The floor number of the unit listed.</td>
<td></td>
</tr>
<tr>
<td>Rooms Count</td>
<td>The number of bedrooms in the listed unit.</td>
<td></td>
</tr>
<tr>
<td>Bathrooms Count</td>
<td>The number of bathrooms in the listed unit.</td>
<td></td>
</tr>
<tr>
<td>Living Rooms Count</td>
<td>The number of living rooms in the listed unit.</td>
<td></td>
</tr>
<tr>
<td>District</td>
<td>Name of the neighborhood where the listed unit is located.</td>
<td></td>
</tr>
<tr>
<td>Sector</td>
<td>Name of the sector where the listed unit is located. Where the sector consists of more than one district.</td>
<td></td>
</tr>
<tr>
<td>Meter Sold Total Area</td>
<td>Area of the unit listed in sqm</td>
<td></td>
</tr>
<tr>
<td>Master Project</td>
<td>Name of the master project that the unit listed belongs to.</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: 'Units View' Dataset Description

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Attribute (variable)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units View</td>
<td>Project Id</td>
<td>Id of the project that the unit listed belongs to.</td>
</tr>
<tr>
<td></td>
<td>Project Name</td>
<td>Name of the project that the unit listed belongs to.</td>
</tr>
<tr>
<td></td>
<td>Unit Number</td>
<td>The number of the listed unit</td>
</tr>
<tr>
<td></td>
<td>View</td>
<td>View of the unit listed (sea view, partial sea view, city view, or internal view).</td>
</tr>
</tbody>
</table>
Table 3: 'Projects Completion Dates' Dataset Description

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Attribute (variable)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects Completion Dates</td>
<td>Project Id</td>
<td>Id of the project.</td>
</tr>
<tr>
<td></td>
<td>Project Name</td>
<td>Name of the project.</td>
</tr>
<tr>
<td></td>
<td>Completion Date</td>
<td>The date when the project was entirely constructed</td>
</tr>
</tbody>
</table>

4.2.3 Data Exploration

In this section, we explored dataset attributes using R programming language to understand the datasets, validate that it satisfies the analytical requirements, and examine the data quality. We started by checking the attribute names in all datasets and ensuring they have no spaces for easier processing. Then the common attribute names between the different datasets were standardized using excel to easily join the datasets in the data processing phase.

- Units Sales Dataset:
  - The dataset consists of 9292 records and 16 attributes. Two duplicate rows were found in the dataset.

![Figure 6: Number of rows and columns for the Units Sales Dataset](image)

- We explored the attributes' data types and found that we have five integer attributes, two numeric or decimal attributes, and nine character or string attributes. One of them was the sale date attribute which should be a date-type attribute, so we converted it to date.

![Figure 7: Data types of Units Sales Dataset attributes](image)
We computed the summary statistics for the numerical attributes such as mean, median, min, max, and quartiles. Also, we explored the number of distinct values, frequencies, and proportions for the categorical attributes.

![Figure 8: Numerical attributes summary statistics for the Units Sale dataset](image)

![Figure 9: Categorical attribute summary statistics for the Units Sales dataset](image)
Figure 10: Sale Date attribute summary statistics

- Nine attributes contained missing values and were handled in the data preparation phase.

**Units View dataset:**
- This dataset consists of 25656 records and four attributes. No duplicate rows were found in the dataset.

<table>
<thead>
<tr>
<th>unitsViews Dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rows: 25656</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number columns: 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of distinct values in the dataset: 25656</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 11: Number of rows and columns in the 'Units View' dataset

- We explored the attributes' data types and found that we have one integer attribute and three character or string attributes.

```
projectId  projectName  unitNumber  view
"integer"  "character"  "integer"  "character"
```

Figure 12: Data Types of Units View dataset attributes.

- We computed the summary statistics for the numerical attributes such as mean, min, max, and quartiles. Also, we explored the number of distinct values, frequencies, and proportions for the categorical attributes.
Figure 13: Summary statistics about the Units View dataset attributes.

- No missing values exist in this dataset.

- **Projects Completion Dates Dataset:**
  - This dataset consists of 138 records and three attributes. No duplicate rows were found in the dataset.

  ```
  projectsCompletionDates Dataset
  Number of rows: 138
  Number columns: 3
  Number of distinct values in the dataset: 138
  ```

  Figure 14: Number of rows and columns in the Projects Completion Dates dataset

  - We explored the attributes' data types and found that we have one integer attribute and two characters or string attributes; one of them was the Completion Date attribute which should be a date type attribute, so we converted it to date.

  ```
  projectId  projectName completionDate
  "integer"    "character"    "character"
  ```

  Figure 15: Data Types of Projects Completion Dates dataset attributes

  - We computed the summary statistics for the numerical attributes, the project Id, such as mean,
min, max, and quartiles. Also, we explored the number of distinct values, frequencies, and proportions for the categorical attributes.

![Table showing summary statistics about the Projects Completion Dates dataset](image)

Figure 16: Summary statistics about the Projects Completion Dates dataset

- No missing values exist in this dataset.

### 4.3 Data Preparation

This phase is critical because the data quality is a crucial factor affecting the model result and the project goal's success. In this phase, we consider multiple data processing steps to prepare the final datasets that can be used in the modeling phase. Then we explore and visualize the final dataset to draw some valuable insights.

#### 4.3.1 Data Preprocessing

Data processing steps include data cleaning, data transformation, feature engineering, and feature selection, as shown in the following steps.

##### 4.3.1.1 Data Cleaning

‘Projects Completion Dates’ and ‘Units View’ datasets were completed and did not require any data cleaning. All the following data cleaning steps were done for the ‘Units Sales’ datasets.

- Removing duplicates values: two duplicate rows were found in the ‘Units Sales’ dataset and removed, so the number of records in the dataset reduced to 9290 records.
- In the unit type attribute, some irrelevant data were removed, which are Shops, offices, and villas,
because my study considers only residential units. Also, additional features affect their prices and are not available in our datasets, like land area for villas. In addition, Shops, offices, and villas form only 0.5%, 1.6%, and 2.2% of the dataset, respectively, which is a very small amount of data that will not be insufficient for machine learning algorithms. After removing these values number of records in the dataset was reduced to 8894

- Missing values were handled as follows:
  
  o Unit type, area, and rooms count missing values were inputted manually by comparing the desired unit with a similar listed unit. Similar units were identified in multiple ways, such as the same unit number on the same project or master project in the nearest floor number, where some projects have a unified plan for all real estate units with the same number on the different floors. Other similar units were identified by having the same unit area or bedrooms count either in the same project or master project.
  
  o Floor Number missing values were inputted based on the unit number, as the unit number format consists of the floor number and the unit id. So, for a three digits unit number, the floor number will be the first digit of the unit number, and for a four digits unit number, the floor number will be the first two digits of the unit number. Ten missing values remained for a non-digit unit’s number and were inputted manually.
  
  o Bathrooms count missing values were inputted based on rooms counts and unit type; studios with one bathroom and apartments with the same numbers on rooms count.

- Some incorrect values were discovered in some attributes and updated as follows:
  
  o Floor number attribute had few decimal values and was corrected.
  
  o Some incorrect entries were discovered in the rooms count and unit area attribute after reviewing the plot in Figure 17 that shows the distribution of rooms count over unit areas for each unit type:

    ▪ There was an extreme unit area that occurred for the studio’s unit type, and after reviewing a similar unit, we found that it is an apartment with one bedroom and two bathrooms.

    ▪ Studios should have 0 room count, and that differentiates it from the one-bedroom apartment, so we updated the rooms count for all studios to 0.

    ▪ There was an extreme unit area occurred for the one-bedroom apartment, and after reviewing this unit and comparing it to similar units, we found that it is a three bedrooms apartment.

    ▪ Few apartments had four rooms count, and after reviewing these units and comparing them to similar units, we found that they are three bedrooms apartments.
63 apartments had 0 rooms count; these units were updated by calculating the mean area for the one, two- and three-bedroom apartments, respectively, for every project of the target units. Then the target unit area was compared with the calculated unit areas to identify the nearest area value. Finally, the closest average area's rooms count was assigned to the target unit.

![Distribution of rooms count over Units Areas for each Unit Type](image1)

Figure 17: Distribution of rooms count over Units Areas for each Unit Type

- Incorrect values in the bathrooms count attribute for apartments unit type were updated to have the same number of rooms count. And all studios were updated to have one bathroom.
- To convert the Floor number to digits in the data transformation step, the ground floor (G) was replaced with 0, and the other non-digit floor numbers were removed.
- Box plots were used to check outliers’ presence in the numeric attributes. Outliers were present in the sale price and unit area attribute and removed using interquartile range (IQR) to enhance the accuracy of the model to be built. Figures 16 and 17 show data distribution before and after removing the outliers.

![Sale Price boxplot](image2)

![Unit Area in sqm boxplot](image3)

Figure 18: boxplots of the sale price and unit area before removing outliers.
4.3.1.2 Feature Engineering

In this step, data sets were joined to produce new features.

- ‘Units Sales’ dataset was left joined with the ‘Units View’ dataset based on the common attributes project id, project name, and unit number, so all records in the ‘Units Sales’ dataset will still exist in the joined dataset, and a new attribute ‘unit view’ will be created. Two hundred forty-nine missing values were present in the unit view attribute, as shown in the summary statistics of the attribute in Figure 20, and these missing values were from the dataset.

![Figure 20: Summary statistics for the ‘view’ attribute in the joined dataset](image)

- ‘Units Sales’ dataset was left joined with the ‘Projects Completion Dates’ dataset based on the common attributes of project id and project name, so all records in the ‘Units Sales’ dataset will still exist in the joined dataset, and a new attribute completion date will be created. Four hundred fourteen missing values were present in the new completion date attribute, as shown in the attribute's summary statistics in Figure 21, which means these projects are still under construction. Then, the completion date was used to create the new attribute, which is the unit age (in days), by subtracting the sale date from the completion date, and for records with missing completion dates, the age was set to 0. Figure 22 shows a summary of statistics of the unit age of the attribute. In the end, the completion date attribute was
removed from the dataset because it would add no value to the study.

![Figure 21: Summary statistics for the 'completion date' attribute in the joined dataset](image)

![Figure 22: Summary statistics for the 'unit age' attribute](image)

4.3.1.3 Feature Selection

The following irrelevant attributes were removed from the dataset before building the machine-learning models. Still, we kept them in the processing phase because some of them helped us in extracting and inputting the missing values of other attributes as described in the data cleaning step:

- **Unit Type Id and Unit Type**: we have two variables representing the unit type: string variable (Unit Type) and integer variable (Unit Type Id), both attributes were removed because the room count attribute gives us the same information, whereas the 0-room count is a studio and other rooms count values are apartments.

- **Project Id, Project Name, Master Project, and Unit Number**: these variables were removed since they are unique identifiers for the unit so they will cause a model overfitting.

- **Unit Use**: this attribute will add no values to the study since all considered units in the dataset are residential.

- **Living Rooms Count**: this attribute will add no value to the study because all studios should have 0 living rooms, and all apartments have one living rooms (16 apartments only had more than one living room which is most likely a wrong data entry).

- **Sector**: is a group of districts, so we chose to use the district attribute because it is more detailed, and the sector was removed as it will add no value and will be a duplicate attribute.

In Addition, Stepwise Regression was used in the modeling step to select the most relevant attributes to build the model.
4.3.1.4 Data Transformation

In this step, we transformed the attributes to a new data type or form that is more appropriate for machine learning algorithms. This step was performed after data visualization to produce meaningful visualizations.

- Converting rooms count, bathrooms count, and floor number attributes from string to integer data type.
- Encoding categorical attributes: unit view, and district, to convert them to numerical attributes that the machine learning algorithms can handle. One hot encoding technique was used where we created a binary dummy variable for each distinct value, excluding one in each attribute. So, we had nine dummy variables for the district and three for the unit view.
- Scaling numerical attributes by transforming their values to have a mean of 0 and a variance of 1, standardizing numerical attributes will prevent machine learning models from assuming that larger values are more significant. Scaling was done by computing the mean and std of the testing dataset and using it to scale the training and testing dataset, so transformation stays consistent across training and testing datasets.

At the end of the data preprocessing step, the final dataset was created with 8355 records and 18 attributes, including the dependent variable, the sale price.

4.3.2 Data Exploration and Visualization

In this phase, the preprocessed and cleaned dataset was used to create visualizations using Tableau and the correlation matrix using R, which helped us uncover some insights to answer our research questions.

4.3.2.1 What are the historical trends for unit prices in Ajman? And what is the effect of Covid-19 on the real estate units’ market in Ajman?

The chart illustrated in Figure 23 demonstrates the seasonality pattern of average sale prices over the years, where the sale prices increased in 2017 and 2019, decreased in 2018 and 2020, and then tended to increase again in 2021. In addition, average sale prices showed an increasing trend throughout 2017 till 2022. Unit sales peaked in 2019, and the lowest unit prices were in 2018. To have a better understanding, the next step we decompose the data to view more details about the sale price trend in each year.
The chart illustrated in Figure 24 demonstrates that the monthly total sale prices throughout 2017 to 2021 and 2022 were excluded from the chart because we don’t have the complete data for this year. We can observe that there is no clear seasonal pattern. Still, in general, we can see an increase in the total sale prices in July and August when it is the summer vacation in UAE, and then the sale prices increase again in September when people return to work and school. In addition, the prices also increase at the end and beginning of each year, which is December and January.

An interesting observation that can be seen is the effect of the Covid-19 pandemic on the total sale prices in the year 2020; where the minimum sale prices were reached in April 2020, when was the beginning of the lockdown in UAE, then increased again to reach the peak in September 2021. In conjunction with the decrease in the total sale prices in 2020 due to the pandemic, the average sale prices also decreased but didn't reach the lowest average sale prices in 2018, as shown in Figure 23.
4.3.2.2 What are the factors affecting real estate units’ prices?

To understand how the number of bedrooms and the location of the unit affecting unit price, we plotted the chart illustrated in Figure 25, which shows that unit prices vary between districts, where ‘Rumaila 2’ district had the highest prices and ‘Humaideya 1’ district had the lowest prices. It’s also showing that the number of bedrooms within the same district affects the unit price positively, so generally, as the number of bedrooms is increased, the unit price increases, except in the ‘Aamra’ district where the price of a two-bedroom apartment is higher than a three-bedroom apartment. In addition to ‘Nuaimeya 1’ where the studio price (0 rooms count) is higher than the 1-bedroom apartment.

Another interesting observation is that the price of units with the same number of bedrooms varies between districts. For example, the one-bedroom apartment in ‘Rumaila 2’ has a higher price than the 2- or 3-bedrooms apartment in other districts like ‘Rashideya 2’. To understand these cases, we investigated more by adding attributes like the unit age and area as shown in the following plots.

In Figure 25, the plotted chart shows the distribution of average unit prices over unit areas (in sqm) per room count and district, and we can see that unit area positively affects unit price,

Unit areas in ‘Rumaila 2’ are the biggest; at the same time, ‘Humaideya 1’ unit areas are the smallest, which is why ‘Rumaila 2’ has the highest unit prices and ‘Humaideya 1’ has the lowest unit prices.

Units’ areas increase with the number of rooms within each district, and the area of a unit with the same number of bedrooms varies between districts, so the unit area is a reason for the difference in unit prices from one area...
to another. To that end, unit areas can't explain the case of 'Aamra' and 'Nuaimeya 1' mentioned above; that's why we checked the unit age in the following plot in Figure 27.

Treemap Plotted in Figure 27 illustrates the relationship between the average unit prices and unit age per district and room count. The color shows the average unit age, the size indicates the average sale price, and marks are labeled by district and room count. We can see that the darker blocks have smaller sizes, which means the units with higher ages have lower prices. This could explain why the prices of studios in 'Nuaimeya 1' are higher than the one-bedroom apartments, as the studios are newer than the one-bedroom apartment. In addition, it could also explain why the prices of two-bedroom apartments in ‘Aamra’ are higher than the prices of three-bedroom apartments, as the two-bedroom apartments are newer than the three-bedroom apartments.
As illustrated in Figure 28, in general, units with a sea view have the highest prices, then the partial sea view, city view, and internal view. But this sequence was not followed in all districts, as shown in Figure 29. So, we used a chart illustrated in Figure 30 that demonstrates the average unit price and unit age per district and unit view where the color shows an average sale price, size indicates the average of the unit area, and labels show the unit age to have a better understanding. We can observe that in ‘Rumaila 2’, units with a city view had higher prices than the units with a partial sea view even though their areas are almost equal. This is because of the difference in unit age, where units with city view were newer than the partial sea view. In ‘Rashideya 1’, the units with a partial view had higher prices than the units with sea view even though the average area for sea view units was higher than partial sea view units, and this again because of the units’ age where sea view units were older than partial sea view units. In ‘Nuaimeya 1’, units with an internal view had higher prices and older ages than units with a city view. This can be justified by unit area, where units with an internal view were bigger than units with a city view. Finally, in ‘Liwara 2’, units with an internal view had the highest prices, then comes the partial sea view, city view, and sea view, respectively, and this also can be justified by unit area, where units with an internal view were the biggest then partial sea view, city view and sea view respectively.
The plot illustrated in Figure 31 shows the average unit prices and unit areas (in sqm) for each floor number. We observed that, in general, units on higher floors were bigger in the area and higher in price than units on lower floors, however, few high floors had small average areas and, therefore, less price.
The heatmap in Figure 32 shows the relationship between the unit prices, bedrooms count (rooms count), and bathrooms count. We observed that the unit prices increased as the bedroom and bathroom count increased. We also observed that units always come with a given number of bathrooms same as the number of bedrooms or more, and the unit price for a given number of bedrooms gets higher by increasing the number of bathrooms.
As we saw in the previous charts and their analysis, all selected features affect unit prices. The correlation plot in Figure 33 shows the direction and strength of the relationship between the sale price and the numerical attributes. Unit Area has the highest correlation with the sale price, followed by rooms count and bathrooms count, which are equally correlated with the sale price. Finally, floor number and unit age are similarly correlated with the sale price, but in different directions, where unit age affects sale price positively; however, floor number affects sale price negatively. The correlation plot also shows a relationship between other attributes, such as the high correlation between unit area, rooms count, and bathrooms count.

![Correlation plot for the numerical attributes](image)

Figure 33: Correlation plot for the numerical attributes

4.4 Modeling

After preparing the final dataset, we built the machine learning models using R programming language by conducting several steps. We initially started by splitting the dataset into training and testing datasets by a split percentage of 80% for training and 20% for testing. The training dataset was used to build the machine learning models and testing dataset for evaluating the models. After that, we scaled the numerical attributes (as mentioned in the data transformation section) by computing the mean and std of the testing dataset and using it to scale the training and testing dataset so transformation stays consistent across training and testing datasets.

The next step was to use a Stepwise Regression model to select the most significant attributes in building the models. Finally, three machine learning models were built, which are Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Gradient Boosting (GB).

Five-fold cross-validation was used in building the models, the training dataset was split into five datasets, and models were built five times. Each time, one of the five datasets will be the testing dataset, and the others will
be training datasets. The average performance measure of the five models is reported as a performance measure of the model (Krishna, 2021). After that, the 20% testing dataset was used as the final evaluation of the machine learning model. Cross-validation allows us to avoid model overfitting of the training dataset and to ensure the robustness of our model in predicting the sale price.

Using suitable parameters for ML models significantly affects the model’s accuracy in predicting the sale prices. Hence, we used Genetic Algorithm (GA) for hyperparameter tuning the SVR and GB models. The root mean square error (RMSE) was used as an accuracy measure for the objective function of the Genetic Algorithm. In addition, $R^2$, RMSE, and MAE were used to evaluate the model’s accuracy.

### 4.4.1 Stepwise Regression for Feature Selection

In this step, a Stepwise Regression was used to select the most significant attributes by building multiple regression models from the 17 predictor attributes by adding and removing attributes based on p-values in a stepwise technique until all candidate variables were added or removed from the model. The resulting Multiple Regression model had the most significant attributes with a p-value less than 0.001.

Figure 34 shows the final output model of the Stepwise Regression process, where we had 15 attributes left out of the 17 predictors attributes. Rooms count and internal view (one of the dummies attributes from the view attribute) attributes were excluded, so they are not significant in predicting the sale price, and this is reasonable based on our visualization and analysis where there was multicollinearity between room count unit area with a correlation coefficient 0.9 (as showing in Figure 33) which is a very high correlation so using one of them will be enough and will produce more reliable models. So, in this case, the unit area was chosen over the room count because it correlates more with the sale price than the bathrooms count.
4.4.2 Multiple Linear Regression

The MLR model was built using the 15 selected independent variables. Five-fold cross-validation was used. All independent variables had a p-value less than 0.001, and the Regression equation was as follows:

\[
\text{Sale Price} = 282407 + 9408 (\text{FloorNumber}) - 17956 (\text{bathroomsCount}) + 160563 (\text{MeterSoldTotalArea}) - 96445 (\text{unitAge}) + 286189 (\text{district AjmanIndustrial1}) + 171649 (\text{district Humaideya1}) + 154597 (\text{district Liwara2}) + 130123 (\text{district Nuaimeya1}) + 137231 (\text{district Nuaimeya3}) + 90245
\]
(district_Rashideya1) + 132404 (district_Rashideya2) + 179389 (district_Rashideya3) + 318127 (district_Rumaila2) + 80382 (partialSeaView) + 85481 (seaView)

Figure 35: The output model for Multiple Linear Regression

The final performance measures for the tested MLR model were as follows:

- \( R^2 \) indicates that 80.61% of the variation in the sale prices in the testing dataset can be explained by the independent variables.
- MAE indicates that the average of the residuals in predicting the sale prices of the testing dataset was 72226.91.
- RMSE indicates the average error value in predicting the sale prices is 96771 AED.
4.4.3 Support Vector Regression

Using the suitable parameters for ML models significantly affects the model accuracy in predicting the sale prices, so we started by setting the parameters of the SVR. The "Radial Basis Function" kernel function was used, the default function in the Support Vector Regression model (Dobilas, 2020). Then, the Genetic Algorithm was used to tune the rest of the parameters of the SVR, which are the penalty factors (cost), the decision boundary gamma Curviness rate (gamma), and the width of the decision boundary epsilon. The population size of the Genetic Algorithm was set to 50 models with different values of hyperparameters, and three generations were run. The Root Mean Square Error (RMSE) was used as an accuracy measure for the objective function of the Genetic Algorithm. The best tune for the SVR was 9.943271, 0.1061853, and 0.02977909, representing the cost, gamma, and epsilon, respectively. The support Vector Regressions were built for the 15 selected variables with five-fold cross-validation and produced 5206 support vectors. Finally, the built model was evaluated with the testing dataset.

--- Genetic Algorithm ---------------
GA settings:
Type = real-valued
Population size = 50
Number of generations = 3
Elitism = 2
Crossover probability = 0.8
Mutation probability = 0.1
Search domain =
    lower 0.1 0.1 0.01
    upper 10.0 10.0 0.10
GA results:
Iterations = 3
Fitness Function value = -71872.34
Solution =
    x1  x2  x3
    [1,] 9.943271 0.1061853 0.02977909
    11660.85 sec elapsed

Figure 37: Genetic Algorithm result for SVR parameters

svm(formula = salePrice ~ ., data = train_selectedVar, cross = 5, cost = 9.943271, gamma = 0.1061853, epsilon = 0.02977909)

Parameters:
  SVM-Type: eps-regression
  SVM-Kernel: radial
  cost = 9.943271
  gamma = 0.1061853
  epsilon = 0.02977909

Number of Support Vectors: 5206

Figure 38: The output model for the SVR
The final performance measures for the tested SVR model were as follows:

- $R^2$ indicates that 89.87% of the variation in the sale prices in the testing dataset can be explained by the independent variables.
- MAE indicates that the average of the residuals in predicting the sale prices of the testing dataset was 45436.33.
- RMSE indicates that the average error value in predicting the sale prices was 71868.54 AED.

| MAE: 45436.33 | MSE: 5165087650 | RMSE: 71868.54 | R-squared: 0.8987112 |

Figure 39: SVR model performance measures results

4.4.4 Gradient Boosting Model

Initially, we started by setting the parameters of the GB model. As we are dealing with a regression problem, we used the squared error as a loss function in the GB model. Then, the Genetic Algorithm was used to tune the rest of the parameters of the GB model, which are the number of trees, the learning rate (shrinkage), and the maximum depth of each tree. The population size of the Genetic Algorithm was set to 50 models with different values of hyperparameters, and three generations were run. The Root Mean Square Error (RMSE) was used as an accuracy measure for the objective function of the Genetic Algorithm. The best tune for the GB model was 389, 0.0984972, and 3, representing the number of trees, the learning rate, and the maximum depth of each tree, respectively. The Gradient Boosting Model was built for the 15 selected variables with five-fold cross-validation, and the built model was evaluated with the testing dataset.
--- Genetic Algorithm ---

GA settings:
Type = real-valued
Population size = 50
Number of generations = 3
Elitism = 2
Crossover probability = 0.8
Mutation probability = 0.1
Search domain = 
   x1  x2  x3
   lower  1 1e-04  1
   upper  512 1e-01  3

GA results:
Iterations = 3
Fitness function value = -72604.65
Solution =
   x1  x2  x3
   [1.] 386.3728 0.0984972 2.643811
541.02 sec elapsed

Figure 40: Genetic Algorithm result for GB model parameters

```
gbm(formula = salePrice ~ ., distribution = "gaussian",
data = train_selectedVar, n.trees = 389, interaction.depth = 3,
shrinkage = 0.0984972, cv.folds = 5)
A gradient boosted model with gaussian loss function.
389 iterations were performed.
The best cross-validation iteration was 389.
There were 15 predictors of which 15 had non-zero influence.
```

Figure 41: The output model for the GB

The unit area was the most significant attribute influencing the sale price prediction in the GB model.

<table>
<thead>
<tr>
<th>var</th>
<th>relinf</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetersSoldTotalArea</td>
<td>47.60621007</td>
</tr>
<tr>
<td>district.Kumalali2</td>
<td>22.18502271</td>
</tr>
<tr>
<td>unitAge</td>
<td>17.11017829</td>
</tr>
<tr>
<td>seaView</td>
<td>6.01072356</td>
</tr>
<tr>
<td>district.Rashideya3</td>
<td>1.89487353</td>
</tr>
<tr>
<td>district.AjmansIndustrial1</td>
<td>1.41189647</td>
</tr>
<tr>
<td>district.Liwa2</td>
<td>1.01148378</td>
</tr>
<tr>
<td>bathroomsCount</td>
<td>0.68299206</td>
</tr>
<tr>
<td>district.Nuaimeyya3</td>
<td>0.54660968</td>
</tr>
<tr>
<td>district.Rashideya1</td>
<td>0.42041865</td>
</tr>
<tr>
<td>FloorNumber</td>
<td>0.38184557</td>
</tr>
<tr>
<td>partialSeaView</td>
<td>0.30794247</td>
</tr>
<tr>
<td>district.Humaldeya1</td>
<td>0.20381206</td>
</tr>
<tr>
<td>district.Nuaimeyya1</td>
<td>0.17355241</td>
</tr>
<tr>
<td>district.Rashideya2</td>
<td>0.04983669</td>
</tr>
</tbody>
</table>

Figure 42: Attributes influencing the GB model

The final performance measures for the tested GB model were as follows:
- $R^2$ indicates that 90.67% of the variation in the sale prices of the testing dataset can be explained by the independent variables.
• MAE indicates that the average of the residuals in predicting the sale prices of the testing dataset was 47365.76.
• RMSE indicates that the average error value in predicting the sale prices was 67717.82 AED.

Figure 43: GB model performance measures results

4.5 Evaluation

This section evaluated and compared the three built models to select the best model for predicting the sale prices. Table 11 summarizes the performance measures of the three built models, and we can see that GB and SVR models were better than MLR in predicting the sale prices. GB and SVR models produced close results, but the GB model was better by fitting 90.67% of the testing dataset compared to 89.87% for SVR, which is represented by the R² value. In addition, the GB model was more accurate in predicting the sale prices, where RMSE for the GB model was 67717.82 AED, and 71868.54 AED for the SVR model, which means the GB model had less errors, and its predictions were closer to the actual values. On the other hand, the MLR model was less variance in fitting the testing dataset and was less accurate in predicting the sale prices than other models.

Table 4: Models performance measures results

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MAE</th>
<th>RMSE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>72226.91</td>
<td>96771</td>
<td>80.61%</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>45436.33</td>
<td>71868.54</td>
<td>89.87%</td>
</tr>
<tr>
<td>Gradient Boosting Model</td>
<td>47365.76</td>
<td>67717.82</td>
<td>90.67%</td>
</tr>
</tbody>
</table>

To visualize the accuracy of the models, we plotted the actual sales prices in the testing dataset vs. the predicted sale prices using the three models in terms of unit area, as it is the most correlated variable with the sale price. And as shown in Figures 42, 43, and 45, the mispredictions were increased in the high sale prices where the
predicted points were very far from the actual points in the MLR model, while the best fitting was in the Gradient Boosting Model.

Figure 44: Actual vs. predicted sale prices using the MLR model.

Figure 45: Actual vs. predicted sale prices using the SVR model
4.6 Deployment

Finally, the result of this project will be shared with the Department of Lands and Real Estate Regulation to be approved and then deployed to the department’s internal system in the staging environment and tested with recent and live data. After that, the model will be deployed to the live system if testing results are approved.
Chapter 5 - Conclusion

5.1 Conclusion

In the project, we collected and prepared datasets from the Department of lands and real estate regulation. The dataset contains information about historical unit sales transactions and unit details. We implemented several data analytical techniques that helped us draw insights about Ajman's real estate unit pieces. We identified the main attribute affecting the unit prices: the unit area. Other factors affecting unit prices were the number of bedrooms, number of bathrooms, unit view, and unit age. In addition, we observed the effect of the Covid-19 pandemic on the unit sale prices.

The dataset was split into training and testing datasets, where the training dataset was used to build the machine learning models with five-fold cross-validation. The testing dataset was used to evaluate the models. Three machine learning models were built to predict the real estate unit prices: Multiple Linear Regression, Support Vector Regression, and Gradient Boosting. The resulting models were evaluated to choose the best model in terms of accuracy and ability to fit the testing dataset. GB and SVR models produced close results, but the GB model overcomes the SVR model by fitting 90.67% of the testing dataset. In addition, the GB model was more accurate in predicting the sale prices, where RMSE for the GB model was 67717.82.

The findings of this project help produce accurate real estate unit valuations on the spot, reducing the time and steps for Ajman Land and Real Estate Regulation Department services and providing a more reliable real estate market.

5.2 Recommendations and Future work

In future work, we recommend analyzing more factors that may affect unit prices, such as nearby establishments, services in the location, and population data. We also recommend extending the study to include all unit types, not only residential apartments, and studios. In addition, a public website can be created for unit price prediction that will help investors, real estate owners, or brokers to evaluate units at any time.
References


