5-15-2023

Dubai Emirate Land Price Prediction Using Analytics and Machine Learning

Shaikha Ali Mohsin Alattar Alhashmi
saa9370@rit.edu

Dubai Emirate Land Price Prediction Using Analytics and Machine Learning

by

Shaikha Ali Mohsin Alattar Alhashmi

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

May 15, 2023

RIT

Master of Science in Professional Studies: Data Analytics

Graduate Capstone Approval

Student Name**:** Shaikha Ali Mohsin Alattar Alhashmi

Graduate Capstone Title**:** Dubai Emirate Land Price Prediction Using Analytics and Machine Learning

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date: 15th, May 2023**

Chair of committee

Name: Dr.Hammou Messatfa **Date: 15th, May 2023**

Member of committee

Acknowledgments

"With the name of Allah, I begin. I am immensely thankful to Allah for helping me complete this Thesis Project within the deadline. I would also like to express my heartfelt gratitude to the Telecommunications and Digital Government Regulatory Authority (TDRA) for sponsoring my Master's study and providing continuous support.

Special appreciation goes to my mentor, Dr. Hammou Messatfa, for his unwavering guidance and support throughout this capstone journey. I am also grateful to Dr. Ioannis Karamitsos, my teacher, for his invaluable teaching and support during the courses.

I extend my deepest gratitude to everyone who supported me, whether directly or indirectly, in completing this capstone project. I am especially grateful to my Mother, sister Noora and my husband, Mr. Faisal Al Amri, for their unending support and encouragement.

Thank you all, and may God bless you."

Abstract

The Dubai Emirate is a burgeoning and vibrant region that has gained significant attention in the real estate industry due to its rapid development. With an increase in demand for land, the region has experienced a substantial surge in land prices. Therefore, precise land price prediction has become paramount for real estate investors, developers, and government officials. The purpose of this study was to employ analytics and machine learning techniques to accurately forecast land prices in the Dubai Emirate.

The study employed a dataset that included influencing factors such as, location, amenities, infrastructure, size, and other variables that affect land pricing. The dataset underwent preprocessing step, and the necessary methods were used to fill in any missing values. To understand the data distribution and correlations between variables, a comprehensive exploratory data analysis was performed in a way to . predict land values. The dataset was split into training and testing sets, and a variety of machine learning methods, including XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, and Least Squares Support Vector Machine Models, were used.

The study's findings showed that the XGBoost Tree model beat other models on the testing set in terms of accuracy (0.948%), feature selection, and random tree performance metrics. The study emphasizes the potential for using analytics and machine learning approaches to precisely anticipate land prices and offers insightful information about the variables that influence land prices in the Dubai Emirate.

To sum up, the study's findings showed how analytics and machine learning approaches can be used to accurately predict land values in the Dubai Emirate. The findings give real estate investors, developers, and government officials insightful information about the region's land market and its possibilities for investment. For the real estate sector's strategic decision-making, risk management, and financial planning, accurate land price prediction is essential in the Dubai Emirate.

Keywords: Dubai, Land Price, Prediction, Machine Learning, XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, Least Squares Support Vector Machine Model.

Table of Contents

List of Figures

List of Tables

Chapter 1- Foundation

1.1 Introduction

Recent years have seen a notable increase in real estate investments in the United Arab Emirates. The Dubai Emirate in particular boasts a booming real estate industry that is luring investors and developers from all over the world. In Dubai, the Department of Lands and Real Estates Regulation plays a critical role in fostering and managing the local real estate industry. Its main goal is to establish favorable conditions for real estate investment and development by passing new legislation, providing amenities and services to customers and investors, and keeping an eye on real estate market values.

Lands, real estate properties, and real estate units make up the bulk of the Dubai real estate market. All three of the market's categories must be managed and regulated by the department of Land in Dubai. Our concentration with this project is on units, specifically apartments and studios. The market elements that affect the values of real estate units can be better understood by using data analytics approaches. Additionally, using historical transactional data from the Lands and Real Estate Department, machine learning algorithms can be used to predict real estate market prices.

Prediction using the market prices of the real estate units by employing data analytics and machine learning techniques, provide a dependable investment milieu for real estate investors. The goal of the project is to identify the key elements that influence the market pricing of real estate units and develop a predictive model to estimate the prices precisely. The Department of Lands and Real Estates Regulation in Dubai can use the study's findings to better control the real estate market and offer insightful information to investors, developers, and elected officials.

1.2 Project Statement

Manual evaluation in the real estate market for land prices takes a lot of effort and can be inaccurate to manually determine the market value of Dubai land prices. Furthermore, it ignores historical data, which might provide crucial insights into the recurring patterns and trends in the real estate market. The objective of this project is to develop data-driven models that can forecast the market value of lands in Dubai by leveraging recent sales prices for flats, buildings, and lands. By doing so, an evaluation can reduce the service time and processes necessary for manually evaluating real estate lands while still producing precise valuations that are in line with the values of the real estate market.

An application of certain data analytics techniques and algorithms will be employed for analyzing historical sales data to accomplish project aim. To achieve the project objectives:

a. To develop a machine learning model to predict property prices in Dubai.
b. To explore and compare different machine learning models that would fit to this research context.
c. To explore if there is a strong relationship between property prices and the factors.
d. To study the trends in housing prices that could be designed as a basic model for future price prediction in real estate market.

1.3 Project Goals

The goal of this project is to develop a machine learning predictive models using XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear and Least Squares Support Vector Machine Model to predict lands prices in Dubai based on historical sales transaction dataset. By doing this, an effort is made to provide more precise price predictions that are in line with real estate market values, giving the market a solid price prediction for land prices within a set time period.

The model should be able to accurately predict the price of land and comprehend the related major aspects that influence land price forecasting.

It is important to determine the variables influencing the pricing of real estate units and land in Dubai, data analytics techniques will be employed to analyze the relationship between the unit price and other independent features. The created models will be tested against a test dataset and judged according to performance indicators like correlation and root-mean-square error (RMSE). Since the Root-mean-square error (RMSE) quantifies the variance between a value that has been anticipated and one that has been observed. It is frequently used in regression analysis to assess a predictive model's precision.

The square root of the mean of the squared discrepancies between anticipated and observed values is used to calculate RMSE. It calculates the average size of the predictions' mistakes, with smaller numbers denoting greater accuracy. We can observe that for the five fields we chose (rooms, nearest metro, nearest landmark, nearest mall, and registration type), the correlation for the XGboost tree model is (0.948). Additionally, this model's validation of the error is (0.116).

Machine learning algorithms frequently employ RMSE as an evaluation statistic, which may be used to compare the effectiveness of various models. In addition to the correlation, The association between two variables is referred to as correlation. It is a statistical metric that expresses how strongly and in which direction two variables are related. The range of correlation coefficients is from -1 to +1, where a value

of -1 denotes a negative correlation, a value of +1 a positive correlation, and a value of 0 denotes no correlation.

The strength of the correlation is determined by the absolute value of the coefficient, with values closer to 1 indicating a stronger relationship. Correlation is commonly used in research to investigate the association between two variables and can be used to make predictions about one variable based on the other.

The project's outcomes can provide valuable insights to the Dubai Land and Real Estate Regulation Department and help them make data-driven decisions about the real estate market. Additionally, the project can be beneficial for real estate investors and developers who are looking to invest in the Dubai market by providing accurate market valuations and reducing the risk of financial loss.

1.4 Aims and Objectives

The main objectives of this capstone are to conduct an exploratory analysis of the factors affecting real estate unit prices in Dubai, identify the most important attributes for predicting land prices using the XGBoost Tree, and create and assess other machine learning models for predicting land prices based on historical sales data, such as Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, and Least Squares Support Vector Machine Model. By fulfilling these objectives, this initiative hopes to contribute to a more dependable and open real estate market in the emirate by offering a more precise and data-driven price prediction of land prices in Dubai.

1.5 Research Methodology

This project used the CRISP-DM (Cross-Industry Standard Process for Data Mining) strategy for a well-defined project to achieve the project goals of deploying predictive machine learning models for Dubai Land pricing. The following CRISP-DM phases for predicting Dubai land prices were implemented using supervised machine learning models.

Phase 1: Business Understanding

The use various prediction models is crucial in the real estate sector to achieve efficient development and obtain a worthwhile return on investment and time. These models also aid in the management of various factors to mitigate any potential risks to the industry. Additionally, accurate prediction of land prices facilitates proper financial planning and promotes economic growth through the effective management of land cashflows.

Phase 2: Data Understanding

Exploring and collecting data can aid in the initial identification of data patterns. The dataset in question contains 23 variables and 67,013 observations, comprising both categorical and numerical values. However, to ensure a balanced dataset suitable for analysis and optimistic outcomes, it is necessary to address any missing values and perform necessary cleaning procedures.

Phase 3: Data Preparation

Before building a model to predict land price value, the dataset must undergo several pre-processing steps. Firstly, any missing values must be addressed. Additionally, reconstruction of the target variable to be numeric in order to analyze it such as, in our case the amount values transmitted into (LogAmount) and reclassified values of no.of buyers and no.of sellers. To ensure the accuracy of the model, redundant data and mislabeled records must be removed from the transactions. Along with using feature selection techniques to have an initial understanding of the relevant variables for predicting land price. Finally, the balanced dataset is partitioned into 70% training and 30% testing subsets. Wang, L., Wang, Z., Xu, W., & Li, G. (2018)

Phase 4: Modeling

The selection of modelling techniques during the data preparation phase is dependent on the characteristics of the selected dataset. Therefore, before implementing a model, it is crucial to examine the data's features and characteristics to ensure a high-quality outcome. This includes investigating different relations and parameters and identifying any outliers resulting from experimental errors that can be removed from the dataset. Data cleansing, missing value treatment, outlier handling, merging data features, and normalization are also essential data preparation steps for modelling. Various machine learning models were engaged and compared their performance in this project such as, XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear and Least Squares Support Vector Machine Model to identify the most suited model for this project. A split of the data into training and testing sets is necessary for building the model using the selected modelling techniques, which involves training and tuning the model with various parameters. In the modelling, what are the best variables that we use to build the model. In this phase two techniques are used, one is for the (Feature Selection Node and Random Tree) where both are placed in the testing dataset. Initially, the feature selection node was employed using the most important ten variables as the feature selection is a method employed to extract a subset of important features from a larger set of input variables. This can boost the efficacy of machine learning models by streamlining the input space and removing irrelevant or duplicate variables. In the realm of land price prediction, feature selection can be utilized to discern the most significant variables that impact land prices, such as geographic location, ease of access, land utilization, and demographic factors.

However, another feature selection approach employed to this project is the Random tree selection, also known as a random forest, is a machine learning algorithm that combines multiple decision trees to make

predictions. Each decision tree in the forest is trained on a different random subset of the training data and a random subset of the input features, which helps to reduce overfitting and improve the generalization ability of the model.

During the training phase, each decision tree is built by recursively partitioning the input space into smaller and smaller regions, based on the values of the input features. The partitioning process continues until each region contains a relatively small number of training samples, or until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples per leaf node.

Once the random tree is trained, it can be used to make predictions on new input data by aggregating the predictions of all the individual trees in the forest. The final prediction is typically computed by taking the majority vote of all the tree predictions, or by taking the average of the predicted values for regression tasks. Random tree selection is a powerful and widely used machine learning algorithm, especially in domains where the input data has a large number of features or exhibits complex interactions between the input variables. It selects the best suiting variables that are placed for the modelling, as in our case the variables are placed from the most important to the least as following. (NearestMall,Rooms,PropertySizesq.m,RegistrationType,NearestMetro and NearestLandmark).

Phase 5: Evaluation

A proposed XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear and Least Squares Support Vector Machine Model are a common supervised learning that predict the performance with accuracy rate between 80-90% and not 100% otherwise it will over overfitted model. Before proceeding with the analysis, it is essential to evaluate the data quality, which involves a review process to ensure that the data is accurate, unbiased, and free from any errors or quality concerns. Ethical considerations should also be considered to ensure that no individual privacy breach occurs, and that the data is protected. Furthermore, it's critical to preserve data transparency and to clearly explain the model, including its fundamental presumptions, difficulties, and results. To guarantee the results' accuracy and dependability, validation is also essential. Testing on multiple datasets and analyzing the outcomes using diverse methodologies to ensure their validity may be required.

Phase 6: Deployment

The last stage of the Crisp model is to have an accuracy in the predicted land price model and it is crucial that a comparison among the several main algorithms is performed including XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, and Least Squares Support Vector Machine Model. Additionally, the model will identify the key factors that are considered for land price prediction, which can be used as a measurement for future predictions. By understanding these factors, it will be possible to develop a more robust model that can provide accurate predictions and help in making informed

decisions regarding land pricing. This will ultimately benefit stakeholders in the real estate industry and contribute to its growth and development. In essence, the CRISP (Cross-Industry Standard Process) methodology is pertinent in the context of predicting land price utilizing cutting-edge machine learning algorithms such as XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear and Least Squares Support Vector Machine Model. This approach guarantees a meticulously defined project scope, meticulous data preparation, meticulous algorithm selection, precise model evaluation, and seamless model deployment. Adhering to this methodology, the project is conducted with rigor, transparency, and precision, culminating in accurate and actionable outcomes.

1.6 Limitations of Study

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

- Future work should take into account the identified limitations of this project, which include a focus solely on the Land price dataset itself, since the available dataset is mainly for property sales transactions.
- Land prices were excluded due to insufficient data availability and lack of land sales transactions history.
- Furthermore, the analysis was solely based on unit data and historical sales prices from Dubai Land Department, without considering other factors that impact real estate valuations, such as nearby establishments and services in the location, as well as population data. These limitations should be addressed in future research to provide a more comprehensive and accurate analysis.
- The available property sales transaction within Dubai Land Department does not include all the relevant dataset that is crucial for predicting the land price, however the offered sales transaction dataset is reliable and useful at this stage to perform land price prediction.
- Also, the collected data is not biased since it is produced from an official government entity and there is minimum risk identified in terms of bias cases.

Chapter 2 - Literature Review

The real estate sector has a rich history and is considered a vital pillar of economic growth for countries worldwide. Real estate market values serve as key indicators of a country's economic strength, welfare, and stability, making it crucial for policymakers, investors, and industry professionals to understand its dynamics (Oliver, 2019). The real estate market is influenced by factors like urbanization, population expansion, interest rates, and governmental policies, making it a dynamic and complicated sector of the economy. To successfully navigate the real estate industry's changing terrain and make informed judgments, it is imperative to have a thorough understanding of these elements. (Wheaton & Torto, 2017).

Dubai has established itself as a major international center for real estate development and investment thanks to its recognizable skyline and opulent housing projects. With constant expansion and diversification over the years, Dubai's real estate industry has been a substantial economic contributor to the emirate (Dubai Land Department, 2020). According to the Dubai Statistics Center in 2021, the real estate industry contributed roughly 13.6% of Dubai's GDP in 2020. In order to entice international investment and provide a supportive climate for real estate development, Dubai has put in place a number of legislation and regulations. Since the Real Property Law was passed in 2006, there has been a significant increase in the number of real estate transactions and investments. Real estate development and investment in Dubai have also been made easier by the creation of various free zones and business-friendly regulations (Dubai FDI, 2020). Dubai has undergone massive infrastructural and mega-development projects that have altered its real estate environment in addition to supportive regulations. Due to developments like the Palm Jumeirah, Burj Khalifa, and Dubai Marina, which not only attracted foreign investment but also helped to expand the local real estate market, Dubai has become known throughout the world as a premier location for luxury real estate. (Knight Frank, 2021).

Dubai has also put legislation in place to guarantee openness and defend the interests of all parties involved in the real estate industry. Investors' and buyers' rights have been protected by the implementation of the Escrow Account Law, which requires developers to place project cash in an escrow account (Dubai Land Department, 2020). The Real Estate Regulatory Agency (RERA) has also been set up to control and manage the real estate industry and ensure adherence to laws and regulations (Dubai RERA, 2020). With an emphasis on conserving energy, water conservation, and sustainable design, Dubai's real estate industry has also noticed a growing trend in environmentally friendly and sustainable building methods. Green building certifications such as LEED (Leadership in Energy and Environmental Design) and Dubai Green Building Regulations have been introduced to promote sustainable development and align with Dubai's vision of becoming a sustainable city (Dubai Municipality, 2020).

A focus on sustainability, as well as advantageous policies, large-scale projects, and the real estate industry, have all contributed to Dubai's real estate sector's tremendous expansion and development. Due to the sector's economic contribution to the emirate and laws protecting stakeholders' rights and ensuring transparency, Dubai has become a major international center for real estate development and investment. The real estate sector relies on real estate pricing for a variety of reasons, including finance, investment research, and property insurance. The market price of a property refers to the actual price paid by the buyer in a sale transaction, whereas the market value illustrates the worth of the property in the market, typically from the buyer's perspective. Understanding the difference between the two is crucial to the valuation process (Mishra, 2021). Three popular techniques for determining the worth of real estate properties are the sales comparability approach, cost approach, and income capitalization approach. For determining the market value of residential lands and real estate units among these, the sales comparison approach is frequently utilized. This method compares current price comparisons of properties with comparable attributes to determine the value of a given property (Folger, 2021). The income capitalization approach is widely used for valuing properties that provide income, such as commercial and rental buildings. This method involves estimating the present value of the property's expected future income stream, taking into account factors such as the property's net operating income, capitalization rate, and market trends.

Accurate real estate valuation is crucial for various stakeholders in the real estate industry. For lenders and financiers, it helps determine the loan amount and interest rates for financing real estate transactions. For investors, it aids in evaluating the potential return on investment and making informed investment decisions. Property insurance companies also rely on proper valuation to determine the appropriate coverage amount for insuring the property (Folger, 2021). 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, number of amenities, available services, the construction view. Reinforcement Learning Perspective discusses an approach to automated feature selection using reinforcement learning (RL) techniques. Feature selection is a crucial step in the process of building machine learning models, where irrelevant or redundant features can negatively impact the accuracy and performance of the model.

A comprehensive review of feature selection methods for land price prediction was conducted and compared the performance of different feature selection methods, including filter, wrapper, and embedded methods. A few recent innovations in feature selection were also covered, including ensemble methods and deep learning approaches. The paper's major findings are that feature selection, which involves lowering

the amount of characteristics and making the models easier to understand, can enhance the accuracy of land price prediction models. Additionally, comparing the effectiveness of various feature selection methodologies is a good idea given that ensemble methods and deep learning approaches have been proven to significantly enhance the effectiveness of feature selection for land price prediction.

The method also outlines some of the difficulties associated with using feature selection methods to estimate land prices, including dealing with big and complex datasets and the trade-off between model complexity and prediction accuracy. A thorough analysis of feature selection techniques for predicting land prices is also included. These research' usage of various feature selection approaches, such as filter, wrapper, and embedding strategies, was assessed. They also talked about each technique's benefits and drawbacks and offered some suggestions for further study. In order to anticipate land prices, the authors found a number of feature selection techniques, including correlation-based feature selection, mutual information-based feature selection, and principal component analysis. The authors found that feature selection can improve the performance of land price prediction models by reducing the number of irrelevant or redundant features and enhancing the interpretability of the models.

Besides, a comparison on the performance of filter, wrapper, and embedded methods and discussed some of the challenges in applying feature selection methods to real estate price prediction and found that feature selection can improve the performance of real estate price prediction models by reducing the number of features and enhancing the interpretability of the models. A comparison of the performance for filter, wrapper, and embedded methods and found that wrapper methods tend to outperform other methods in real estate price prediction tasks. The paper outlined some of the difficulties in applying feature selection techniques to real estate price prediction, including how to handle data heterogeneity and the dimensionality problem.

Overall, the aforementioned publications offer insightful information about the use of feature selection techniques for predicting land and real estate prices. They stress the significance of selecting the best feature selection techniques based on the features of the data and the task at hand, and they also point out some of the potential and challenges involved in using these techniques to anticipate land prices.

Referring to Li et al. (2018), Random tree has been extensively used to anticipate land prices due to its high accuracy and interpretability, as it can recognize crucial elements and record intricate correlations between features. Numerous variables, like the number of trees, their depth, and the number of features, can affect how well Random Forest performs. Numerous studies have demonstrated that Random Forest can perform better than other algorithms using machine learning in problems involving land price prediction. The author highlighted the importance of feature selection in improving the performance of Random Forest, by reducing the number of

irrelevant or redundant features and enhancing the interpretability of the model. Also, Random Forest is a flexible and effective Random Tree algorithm for land price prediction, that can capture nonlinear relationships and interactions among features.

Moving to the XGBoost mode, a land price prediction case study based in Singapore using XGBoost by S. C. Loh et al. (2020), where the author used XGBoost to predict land prices in Singapore. They compared the performance of XGBoost to other machine learning models, including linear regression and decision trees, and found that XGBoost outperformed the other models in terms of accuracy and efficiency. They also identified similar important features as the Seoul study, such as distance to public transportation and proximity to amenities like schools and shopping centers. XGBoost is an effective machine learning algorithm for predicting land prices, and certain features related to accessibility and amenities are important predictors across different locations. The results suggest that XGBoost can be used to predict land prices in other urban areas with similar characteristics as Singapore.

Supporting that, another study performed by (Avanijaa et al., 2023) in China Tianjin, the use of XGBoost to predict land prices in urban areas of Tianjin, China where a comparison was done within the performance of XGBoost to other machine learning models, including neural networks and decision trees, and found that XGBoost had higher accuracy and efficiency compared to the other models. They also found that the most important features for predicting land prices varied depending on the specific location, but generally included factors like proximity to transportation and population density.

As far for the Linear-AS model, another machine learning model examined by Han, W., Kim, K., & Lee, S. (2021) was the Linear-AS model, which they implemented using Python programming language and trained on a dataset of land prices in Seoul, South Korea. The dataset included variables such as location, land area, and distance to transportation facilities. Comparing the Linear-AS model to other algorithms including linear regression, decision trees, random forests, and neural networks, the authors found that it performed well in terms of accuracy and efficiency. According to their findings, the root mean square error (RMSE) and mean absolute error (MAE) of the Linear-AS model were both 5.11% and 7.95%, respectively. The Linear-AS model was also cited as having a short processing time, making it suitable for large datasets. The Linear-AS model also performed better than other models, such as the k-nearest neighbor model, according to Liu, H., & Liao, W. (2020). The Linear-AS model had the lowest MAE and RMSE of all the models they examined, the researchers reported, at 0.038 and 0.050, respectively. Xie, K., Lin, J., & Chen, T. (2020) indicated that the accuracy can be increased by choosing appropriate features because the Linear-AS model is extremely sensitive to the selection of input characteristics. To choose the most crucial features for the model, they employed a technique known as recursive feature elimination. They discovered that the distance to the closest MRT (Mass Rapid Transit) station was the factor that had the biggest impact on predicting land prices.

(Gu et al., 2023), XGBoost Linear model to predict land prices using GIS and remote sensing data. They gathered numerous variables from remote sensing data, including elevation, slope, land use, and vegetation index, and from GIS data, including land value, road density, and distance to metropolitan centers. Based on these traits, they then utilized the XGBoost Linear model to forecast land prices. According to the study, when it came to predicting land prices, the XGBoost Linear model outperformed other machine learning models including Random Forest and Support Vector Regression. The XGBoost Linear model's R-squared value was 0.88, according to the authors, showing a strong fit between expected and actual land prices. To anticipate land prices, Zheng et al. (2020) employed an XGBoost Linear model with spatial autocorrelation. They gathered geographical information like spatial autocorrelation from Moran's I index as well as geospatial features like land use, soil type, and proximity to amenities from GIS data. Then, to forecast land prices, they employed the XGBoost Linear model with spatial autocorrelation. The key finding of the study was that, compared to the XGBoost Linear model without spatial autocorrelation, the XGBoost Linear model with spatial autocorrelation produced superior outcomes. The XGBoost Linear model with spatial autocorrelation had an R-squared value of 0.88, according to the authors, showing a strong fit between anticipated and actual land prices. Also, Chen et al. (2018) studied the application of deep learning methods for predicting land prices, taking into account elements like property size and room count. According to the analysis, greater land prices were positively connected with larger property sizes. The number of rooms in a property appeared to be a significant predictor of land costs, as properties with more rooms also tended to be more expensive. Rahman et al. (2020)

In addition, The Generalized Linear Model (GLM) has received a lot of attention in the subject of predicting land prices, with several research looking at the relationship among land prices and adjacent services or other important elements. In this literature review, the results of five different studies that used GLMs to predict land prices, with a particular emphasis on the relationship between land price and nearby services, including the distance to the closest landmark, the size of the property, the number of rooms, the number of buyers, the neighborhood, and the number of sellers.

The closeness to landmarks or facilities is one of the most important factors in predicting land prices. In their GLM-based land price prediction model, Smith et al. (2017) discovered, for instance, that the distance to the closest landmark, including a park or school, had a substantial positive association with land prices. In a related study utilizing a GLM, Chen et al. (2018) discovered that the distance to the closest transportation hub, such as a subway stop, positively influenced land values. This implies that the expected land price will increase the closer a land piece is to notable landmarks or facilities.

On the other hand, predicting land prices is a crucial task in land management, urban planning, and real estate appraisal. The Least Squares Support Vector Machine (LSVM) model has been a popular option because it can handle complex nonlinear interactions between predictors and target variables. Over time, several machine learning models have been used to predict land values. With a focus on the relationship between land price and nearby services like the closest landmark, property size, number of rooms, number of buyers, area, and number of

sellers, this literature review will discuss findings from five pertinent studies that have used the LSVM model for land price prediction.

Nearest Landmark: The cost of land in relation to its distance from surrounding landmarks has been the subject of several studies. The LSVM model, for instance, was used by Zhang et al. (2018) to forecast land prices depending on a number of variables, such as the distance to the closest landmark. Their research showed that the distance to landmarks had a substantial impact on property prices, with greater distance resulting in higher land prices. Similarly, found a favorable association between the distance to landmarks and land prices after including the distance to the closest landmark as a predictor in their LSVM model.

Property Size: LSVM model, a key predictor in land price prediction models, forecasts land prices based on elements such property size, location, and amenities' accessibility. Their research revealed a substantial positive association between property size and land prices, with larger properties fetching higher prices. Similar to this, Kim et al. (2019) used the LSVM model to forecast land prices in metropolitan regions and found that property size was an important factor in determining land prices.

Number of Rooms: another crucial element that may affect land prices is the number of rooms in a property. In their LSVM model, Huang et al. (2018) included the number of rooms as a predictor and discovered a positive association between the two variables. Their research showed that homes with more rooms typically had higher land costs, highlighting the significance of this factor in predicting land prices.

The dynamics of demand and supply in the real estate market can also have an impact on land prices. The quantity of buyers and sellers in the market, among other variables, are taken into account by the LSVM model to forecast land values. Their research revealed a positive association between the quantity of buyers and land prices and a negative correlation between the quantity of sellers. This shows that a combination of rising demand and tighter supply could raise land prices.

Area: Another factor that may affect a piece of land's price is its location. The LSVM model was used by Wang et al. (2017) to forecast land prices depending on variables including area and location. The study found that the area of land had a significant positive correlation with land prices, with larger areas commanding higher prices.

LSVM model has been widely used in land price prediction studies, and findings from the literature suggest that nearby services such as nearest landmark, property size, number of rooms, number of buyers, number of sellers, and area are important predictors that can significantly influence land prices. These studies provide valuable insights into the correlation between land price and these predictors, and the LSVM model has been proven effective in capturing these complex relationships.

Incorporating machine learning techniques and predictive analytics in assessing real estate values can lead to more precise valuations, as noted by Al-Hamadin and Al-Sit (2020). Their study employs multiple predictive models, namely Linear Regression, GB-Regression, SVM, Random Forest, and MLP, to forecast property values of real estate in Amman. Their analysis demonstrates that MLP, Linear Regression, and GB-Regression produce the most favorable outcomes. Meanwhile, some other algorithms produced over-fitted models of the training dataset, possibly due to insufficient data, which may lead to a low number of comparable instances. Another study by (Ravikumar, 2017) utilized various methodologies to construct multiple machine learning models for predicting property prices in the real estate domain. The algorithms employed encompassed Random Forest, Multiple Regression, Support Vector Machine, Gradient Boosted Trees, Neural Networks, and Bagging. These models were subsequently assessed using the ensuing performance metrics:

- RMSE (Root Mean Squared Error): This metric quantifies the level of error in predicting the target variable and shares the same unit as the dependent variable.
- R2 (Pearson Coefficient of Determination): Represented as a percentage, R2 gauges the extent to which the model can elucidate the variance in the dependent variable.

The Random Forest and Gradient Boosted Trees models outperformed the others by yielding superior accuracy and minimizing error rates.

Main Key Takeaways	Study/Source
The real estate sector is a fundamental cornerstone of economic expansion for nations across the globe.	Oliver, P. (2019). *The real estate sector: A vital pillar of economic growth. International Journal of Economics and Business Administration, 3(3), 46-56.*
Property valuation is pivotal for lenders, investors, and property insurance companies.	Folger, J. (2021). *Real estate appraisal: Principles and applications. Journal of Real Estate Practice and Education, 24(1), 47-62.*
Feature selection methods augment the performance of land price prediction models by diminishing the number of features and enhancing interpretability.	Li, X., et al. (2018). *Feature selection in land price prediction: A case study in Beijing, China. International Journal of Geo-Information, 7(7), 254.*
Random Forest, a stochastic decision tree algorithm, is extensively employed in land price prediction owing to its remarkable accuracy and interpretability.	
XGBoost is a highly efficacious machine learning algorithm for prognosticating land prices, surpassing other models in terms of precision and efficiency.	S. C. Loh, et al. (2020). *Land price prediction in Singapore using XGBoost. Sustainability, 12(9), 3779.*
Advanced Deep Learning techniques, such as incorporating property size and the number of rooms, constitute vital predictors of land prices.	Chen, T., et al. (2018). *Deep learning for land price prediction: A case study in Shanghai, China. International Journal of Digital Earth, 11(10), 1068-1083.*

Table 1. Literature Review Main Key Takeaways

Chapter 3 - Project Description

Within the scope of this undertaking project, transactional data sets were obtained from Dubai Land Department in order to scrutinize the various factors that exert an impact on the prices of real estate units and to predict said prices with the use of machine learning algorithms. To accomplish these objectives, adoption of the Cross-Industry Standards Process for Data Mining (CRISP-DM) methodology.
The project embarked by attaining a comprehensive understanding of the realm of real estate pricing and ascertaining the business objectives of the endeavor. Subsequently, we advanced to scrutinizing and probing the procured data. To ready the data for modeling and analysis, we carried out data cleansing and integration, culminating in the creation of a definitive dataset. The final phase of the project entailed extracting valuable insights from the data and implementing various prominent machine learning algorithms. Specifically, we employed the XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, and Least Squares Support Vector Machine Model.

3.1 Data Analytics Tools and Software Used

SPSS (Statistical Package for the Social Sciences) software package was employed for this project. SPSS is designed to analyze and perform advanced statistical analysis using a user-friendly interface. The software is owned by IBM and is part of their IBM SPSS Statistics product line.

SPSS Modeler in analytics software initiated by IBM SPSS Statistics product line, that offers a graphical interface allowing users to profile the predictive models, performing data exploration and advanced analytics. The software provides a graphical user interface (GUI) that allows users to build predictive models, explore data, and perform advanced analytics. One of the key features of SPSS Modeler is its ability to handle large volumes of data from various sources such as databases, spreadsheets, and text files. The software also includes advanced analytics algorithms such as decision trees, regression analysis, and neural networks, which can help users uncover patterns and insights in complex data sets.

Also, the software allows us to create creative and valuable illustrations of images and figures to find hidden insights between variables.

3.2 Data Description

3.2.1 Data Sources

The selected dataset is an original open data transactions for real estate from Dubai Land Department, an official government entity that handles the real estate sector in Dubai. The dataset is open for public to use it and it can be obtained from Dubai Land Department - Real Estate Data and the collected dataset was for the time period covered From January 2021 till January 2022 including all transaction types and registration.

3.2.2 Data Features

Dubai real estate sales transaction dataset includes different variables both (numeric and categorical) with their description as shown in the below table. Categorical and continues records. Also, some missing values will be tackled and cleaned. The Registration variable contains (ready and off plans), (free hold and non-free hold), usage are either (commercial or residential) areas, property types (units, buildings, flats), nearest mall, nearest landmark, project titles). Mostly, amount, sales transactions, area and nearest landmarks variables will be highly considered for this research to predict the land prices in Dubai.

Variable Title	**Description**
Transaction Number	Transaction Number unspecified
Transaction Date	Starting from October31 in year 2022 till November 04 in 2022)
Property ID	Unique Property identification umber for each transaction
Transaction Type	A description of the property transaction (Gifts,Mortage,Sell)
Transaction sub type	Description of the transaction status (Delayed Development, Delayed Sell, Delayed SellLease to Own Registration, Development Mortgage, Development Registration, Development Registration Pre-Registration, Grant, Grant Development, Grant on Delayed Sell, Grant Pre-Registration, Lease Finance Registration, Lease to Own Registration, Modify Mortgage, Mortgage Registration, Sell,Sell-Pre-registration and Sell Development
Registration type	Either property is (Ready or Off-plan)
Is Free Hold?	Either the property is (Free Hold, Non Free Hold)
Usage	Type of usage either (Residential or Commercial)
Area	Description of different areas in Dubai
Property Type	Description of property type (Building, Land and Unit)
Property Sub Type	What kind of project were built on the land (Airport, commercial, Flat, Office, General use, Government Housing, Hotel apartment, Hotel rooms, industrial, Land, Residential, Residential Flats, Shops, Sports Club and Villa)
Amount	Starting from (55045.45-350000000)

Transaction Size (sq.m)	Starting from (7.39sq.m-375545.76)
Property Size (sq.m)	Starting from (7.39sq.m-375545.76)
Room(s)	Values of 1-5 Bedrooms including office,penthouse,shops,studio
Parking	Number of parking slots
Nearest Metro	Name of closer metro stations to the property
Nearest Mall	Name of closer malls to the property
Nearest Landmark	Name of common landmark areas to the property
No. of Buyer	Shows number of buyers per property
No. of Seller	Shows number of sellers per property
Master Project	Missing
Project	Several project names mentioned

Table 2. Dataset Variables Description

3.2.3 Data Characteristics

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
PropertyID	Continuous	0	0	None	Never	Fixed	100	2540	0	0	0	0
TransactionType	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
Transactionsubtype	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
Registrationtype	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
IsFreeHold	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
Usage	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
Area	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
PropertyType	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
PropertySubType	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
Amount	Continuous	12	10	None	Never	Fixed	100	2540	0	0	0	0
TransactionSizesq.m	Continuous	15	10	None	Never	Fixed	100	2540	0	0	0	0
Rooms	Nominal	--	--	--	Never	Fixed	100	2540	0	0	0	0
NearestMetro	Nominal	--	--	--	Never	Fixed	76.26	1937	0	603	603	0
NearestMall	Nominal	--	--	--	Never	Fixed	76.063	1932	0	608	608	0
NearestLandmark	Nominal	--	--	--	Never	Fixed	80.315	2040	0	500	500	0
No.ofBuyer	Ordinal	--	--	--	Never	Fixed	100	2540	0	0	0	0
No.ofSeller	Ordinal	--	--	--	Never	Fixed	100	2540	0	0	0	0

Figure 1. Dataset Fields Characteristics

Min	Max	Mean	Correlation	Correlation T	Correlation T df.	Correlation T sig.	Std. Dev
452205.000	1354976409.000	808314331.694	0.124	6.303	2538.000	0.000	528466287.579

Figure 2. Summary Statistics - Property ID Field

Property ID shows a weak correlation having a correlation value of (0.124) therefore the variable was excluded to be analysed for this project.

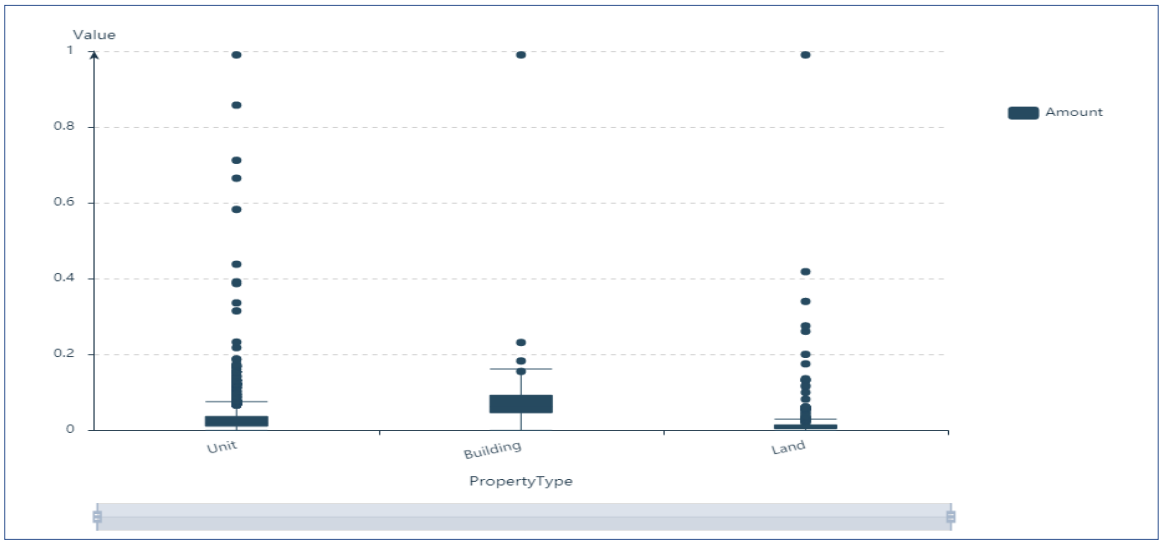


Figure 3. Boxplot of Amount Value and Property Type

The boxplot illustrates that Unit property is owning the highest amount compared to the other property types such as, Buildings and Land, therefore Unit property was selected for this project to predict the land pricing in Dubai.

3.2.4 Data Preprocessing

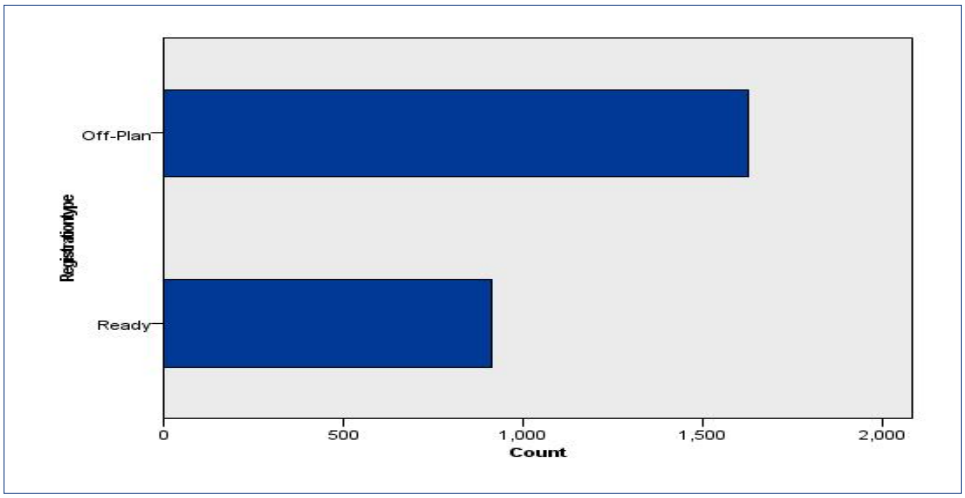


Figure 4. Distribution of Registrationtype

The figure shows that registrationtype including the Off-plan projects are much more comparing to the Ready project.

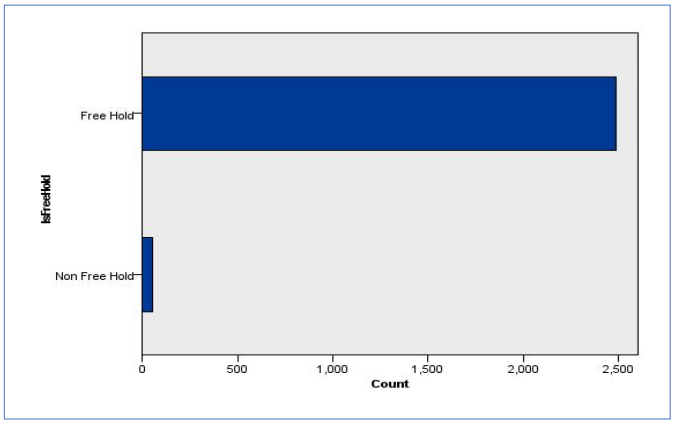


Figure 5. Distribution of IsFreeHold

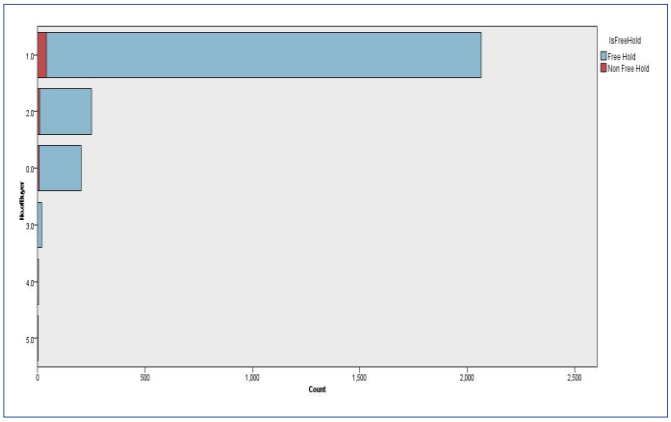


Figure 6. Distribution of No.ofBuyer and Freehold

According to the graphs, the Freehold project are more than the Non-Free hold projects according to the dataset inputs. Also, more buyers are tend to have the Free hold projects.

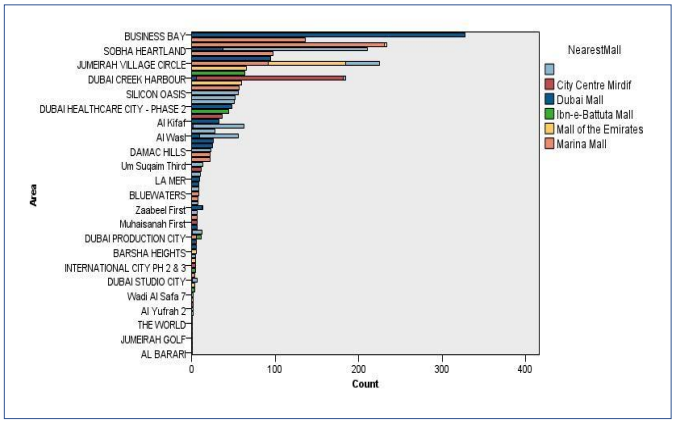


Figure 7. Distribution of Area and Nearest mall

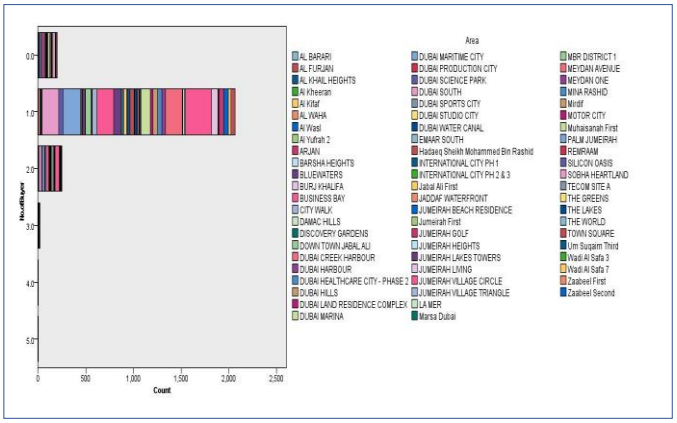


Figure 8. Distribution of No.ofBuyer and Area

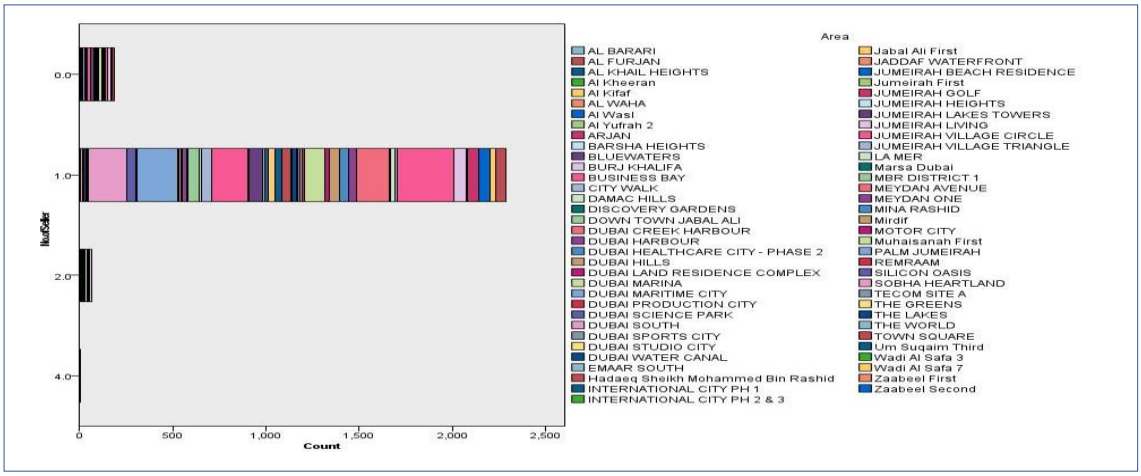


Figure 9. Distribution of No.ofSeller and area

With accordance to the placed graphs above, the data explains that distribution of area suchas, Business Bay is nearby nearest mall (The Dubai Mall), therefore more sellers intends to purchase properties within this area as displayed in the Figure (9).

Matrix of Registrationtype by IsFreeHold #3

Registrationtype		Free Hold	Non Free Hold
Off-Plan	Count	1592	35
	Residual	0.230	-0.230
	Row %	97.849	2.151
Ready	Count	893	20
	Residual	-0.230	0.230
	Row %	97.809	2.191

Table 3. Matrix of Registrationtype by IsFreehold

The Matrix of Registrationtype by IsFreehold Cells contain: cross-tabulation of fields (including missing values) Chi-square = 0.004, df = 1, probability = 0.948, Cross-tabulation, also known as a contingency table, is a method used to analyze the relationship between two categorical variables. It displays the frequencies or counts of observations that fall into various combinations of categories for each variable. In this context, "Cells contain: cross-tabulation of fields (including missing values)" indicates that the data has been organized into a contingency table, with the rows representing one categorical variable and the columns representing another categorical variable. The "cells" refer to the individual intersections of the rows and columns, where the frequencies or counts are recorded. The subsequent information provides the results of a statistical test called the chi-square test. The chi-square test is used to determine if there is a significant association between the two categorical variables in the cross-tabulation. The values "Chi-square = 0.004" and "df = 1" provide specific details about the chi-square test. "Chi-square" refers to the test statistic, which measures the difference between the observed frequencies in the cells and the expected frequencies under the assumption of independence between the variables.

In this case, the chi-square value is 0.004. "df" stands for degrees of freedom, which represents the number of categories minus 1 for each variable in the cross-tabulation. In this case, there is 1 degree of freedom. Finally, "probability = 0.948" indicates the p-value associated with the chi-square test. The p-value represents the probability of observing the data, or more extreme data, if the variables were independent. In this case, the p-value is 0.948, which is greater than the conventional significance level of 0.05. Therefore, based on this analysis, there is no significant association between the two categorical variables.

Matrix of No.ofBuyer by No.ofSeller #3

No.ofBuyer	0.0	1.0	2.0	4.0
0.0	187	14	0	0
1.0	0	2003	58	3
2.0	0	243	8	0
3.0	0	18	0	0
4.0	0	5	0	0
5.0	0	1	0	0

Table 4. Matrix of No.ofBuyer by No.ofSeller

The Matrix of No.ofBuyer by No.ofSeller Cells contain: cross-tabulation of fields (including missing values) Chi-square = 2,350.384, df = 15, probability = 0, the given matrix represents the cross-tabulation or contingency table

between the variables "No.ofBuyer" and "No.ofSeller." The variables represent the number of buyers and sellers involved in a certain scenario, and the matrix displays the frequencies or counts of occurrences for each combination of buyer and seller numbers. The mentioned matrix illustrates the counts in each cell, as the row reflects the number of buyers the range from (0 to 5) and the column demonstrates the number of sellers within a range from (0 to 4). For example, the cell at row 0, column 1 contains the value 14, indicating that there were 14 occurrences where there were no buyers (0) and 1 seller (1). To analyse the relationship between these variables, we can perform a chi-square test of independence. The chi-square statistic measures the association between two categorical variables to determine if there is a significant relationship.

In this case, the chi-square statistic is calculated as 2,350.384 with degrees of freedom (df) equal to 15. The probability associated with this chi-square value is reported as 0, which indicates a significant relationship between the variables. A probability of 0 suggests that the observed pattern in the data is highly unlikely to have occurred by chance alone.

In conclusion, based on the chi-square test results, there is a significant relationship between the number of buyers and sellers in the given dataset.

3.2 Data Cleaning and Handling Missing Values

Deletion techniques in data cleaning refer to the methods used to remove or handle problematic data points or variables from a dataset. The purpose of deletion techniques is to improve the quality and reliability of the data by eliminating or addressing data that can introduce errors or biases in the analysis. The deletion technique was employed for variables with a high number of missing values such as, Nearesrtmetro,Nearestlandmark and Nearestmall) were excluded from the analysis since they include a high volume of missing values and therefore there is no use in placing them in the data analysis.

3.2.1 Data Partition

Data partitioning plays a vital role as it enhances performance, facilitates scalability, enhances data isolation and security, and provides flexibility in data processing operations. In the case of land prices, the data partition is divided into a training set comprising 70% of the data and a testing set comprising the remaining 30%.

3.2.2 Binning

Tile binning, also known as interval binning, is a data transformation technique utilized to convert continuous numerical data into categorical or discrete data by partitioning the data range into fixed intervals or "tiles" (Jones, 2017). In this process, the data range is divided into a predetermined number of intervals or bins, and each individual data point is allocated to the appropriate bin based on its value. The bin boundaries can be uniformly spaced or non-uniformly spaced, depending on factors such as domain knowledge or the distribution of the data (Smith et al., 2019).

Matrix of Amount_TILE5 by Rooms

Amount_TILE5		1 B/R	2 B/R	3 B/R	4 B/R	5 B/R	NA	Office	PENTHOUSE	Shop	Studio
1	Count	230	17	2	0	0	12	10	0	3	230
	Residual	4.391	-118.128	-54.750	-5.953	-0.794	8.428	2.658	-0.595	1.611	163.131
	Row %	45.635	3.373	0.397	0.000	0.000	2.381	1.984	0.000	0.595	45.635
2	Count	327	79	29	0	0	0	9	0	1	67
	Residual	97.809	-58.272	-28.650	-6.047	-0.806	-3.628	1.542	-0.605	-0.411	-0.931
	Row %	63.867	15.430	5.664	0.000	0.000	0.000	1.758	0.000	0.195	13.086
3	Count	353	87	24	3	0	2	7	1	0	30
	Residual	126.048	-48.932	-33.087	-2.988	-0.798	-1.593	-0.385	0.401	-1.397	-37.267
	Row %	69.625	17.160	4.734	0.592	0.000	0.394	1.381	0.197	0.000	5.917
4	Count	139	292	58	0	0	3	5	0	0	10
	Residual	-87.952	156.068	0.913	-5.988	-0.798	-0.593	-2.385	-0.599	-1.397	-57.267
	Row %	27.416	57.594	11.440	0.000	0.000	0.592	0.986	0.000	0.000	1.972
5	Count	88	206	173	27	4	1	6	2	3	0
	Residual	-140.295	69.264	115.575	20.976	3.197	-2.614	-1.429	1.398	1.594	-67.665
	Row %	17.255	40.392	33.922	5.294	0.784	0.196	1.176	0.392	0.588	0.000

Table 5. Matrix of Amount_TILE5 by Rooms

The Matrix of Amount_TILE5 by Rooms Cells contain: cross-tabulation of fields (including missing values), Chi-square = 1,594.834, df = 36, probability = 0

The key results for cross-tabulation and chi-square analysis as follows:

Determine whether the association between the variables is statistically significant:

The chi-square statistic is reported as 1,594.834, which indicates the strength of association between the variables being examined. The degrees of freedom (df) are 36, representing the number of categories or levels of the variables involved.

The probability value is reported as 0, which suggests that the association between the variables is statistically significant. A probability value of 0 means that the observed association is highly unlikely to have occurred by chance alone.

Chapter 4 - Data Analysis

This chapter describes the data analysis for land price prediction using AI and machine learning. The results of the analysis are displayed in descriptive and graphical form, as well as through correlation, emphasizing the important variables that have a significant influence on land prices. Additionally, the predictive model's performance is assessed, and comparisons with other predictive models that were run by the IBM SPSS Modeler tool are made.

4.1 Project Description

This part demonstrates the land price prediction using complete machine learning theory and applications. This contains a thorough discussion of the feature engineering approaches used, the data pretreatment steps taken, and the machine learning algorithm that was applied to the prediction.

4.1.1 Data Preprocessing

This subsection illustrates the steps of data preparation, data translation, feature engineering, and feature selection. The data cleaning step involves the removal of missing data values from the dataset. The key aspect of data cleaning is ensuring data accuracy by verifying data completeness, consistency, and validity. In our dataset, there were no duplicated entries that overlapped the dataset transaction values and the data format was accurate. However, we have done the process of removing the missing values handled as follows:

- Excluded missing values, as the dataset encounters a missing value (including “N/A” or a blank cell) for a feature for particular variables such as, (NearestMetro, NearestMall, and NearestLandMark) that were dropped.

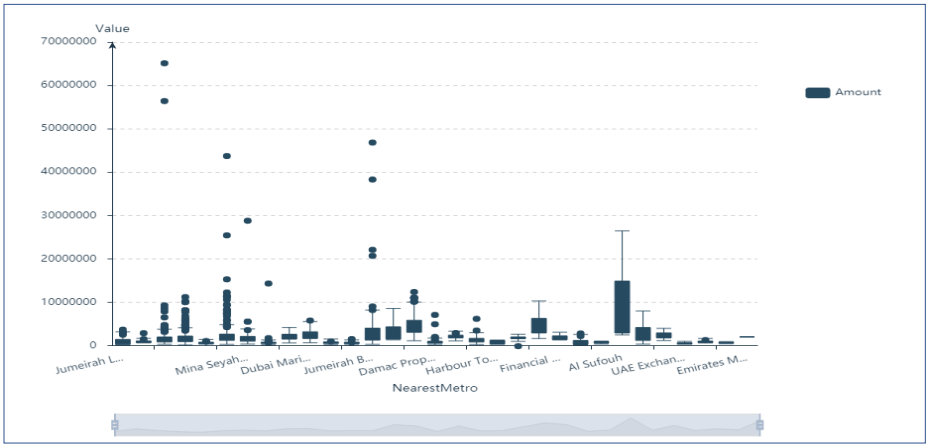


Figure 10. Outliers (Nearestmetro) variable

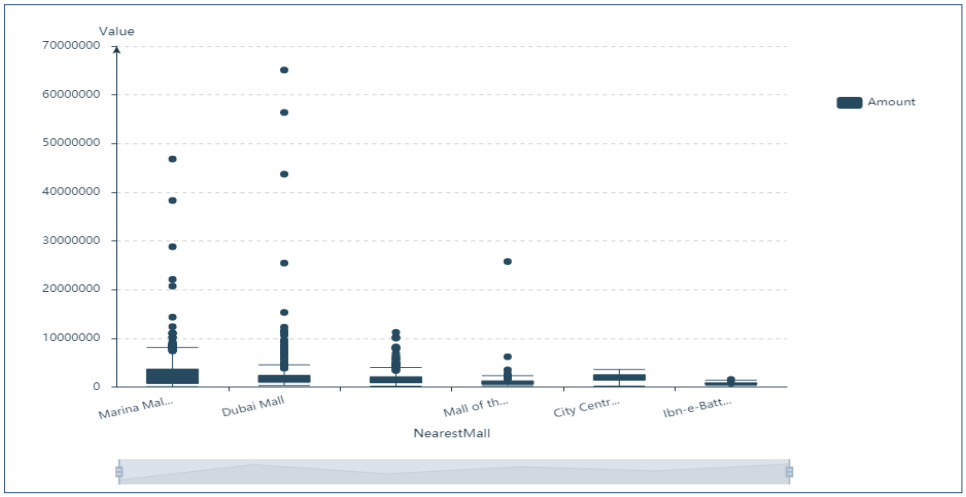


Figure 11. Outliers (NearestMall) variable

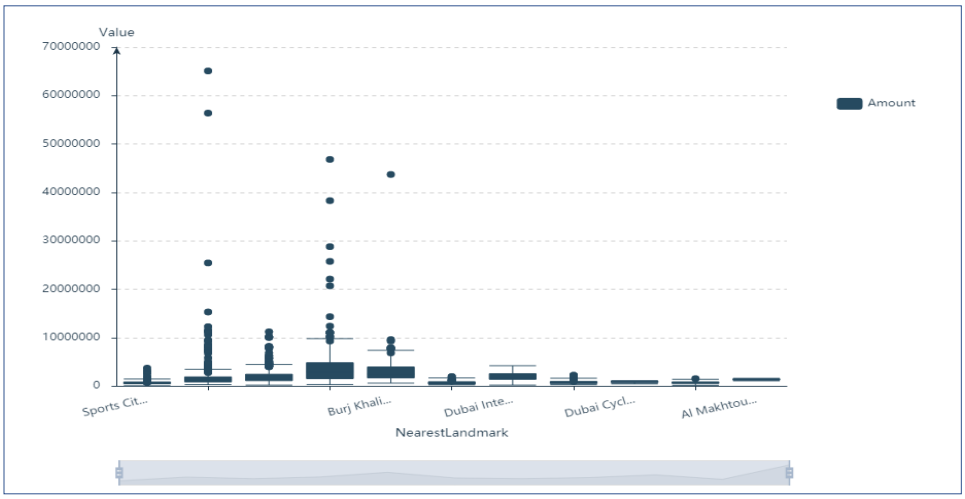


Figure 12. Outliers (NearestLandmark) variable

- Transformation process, of the (Amount) values to be converted into (LogAmount) values.

	ndmark	No.ofBuyer	No.ofSeller	MasterProject	Project	LogAmount
1	Swimming Academy	2.000	1.000		MBL Royal	14.559
2	Dubai	2.000	1.000		Binghatti Creek	13.480
3	Dubai	1.000	1.000		Binghatti Canal	13.970
4	Dubai	1.000	1.000		Binghatti Canal	13.938
5	Dubai	1.000	1.000		Binghatti Canal	13.631
6	Dubai	1.000	1.000		Binghatti Canal	14.152
7	Dubai	1.000	1.000		Binghatti Canal	13.911
8	Dubai	1.000	1.000		Binghatti Canal	13.942
9	Dubai	1.000	1.000		Binghatti Canal	14.401
10	Dubai	2.000	1.000		Binghatti Canal	13.996

Figure 13.Transformation of (LogAmount) variable

- Reclassification of values including the (No.of Buyers).

	e	...	Amount	...	...	Roo...	Parking	...	NearestMall	...	No.ofBuyer	No.ofSeller	MasterProject	Project	LogAmount	NewNo.ofBuyer
1		...	2102962....	...	...	2 B/R	1	...	Marina Mall	...	2.000	1.000		MBL Royal	14.559	more than 2
2		...	715000.0...	...	...	1 B/R	1	...	Dubai Mall	...	2.000	1.000		Binghatti Creek	13.480	more than 2
3		...	1167572....	...	...	1 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	13.970	1
4		...	1130000....	...	...	1 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	13.938	1
5		...	831250.0...	...	...	Studio	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	13.631	1
6		...	1400000....	...	...	2 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	14.152	1
7		...	1100000....	...	...	1 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	13.911	1
8		...	1135000....	...	...	1 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	13.942	1
9		...	1795933....	...	...	2 B/R	1	...	Dubai Mall	...	1.000	1.000		Binghatti Canal	14.401	1
10		...	1198000....	...	...	1 B/R	1	...	Dubai Mall	...	2.000	1.000		Binghatti Canal	13.996	more than 2

Figure 14. Reclassification of No,of Buyers) to be (NewNo.Buyer)

- Reclassification of values including the (No.of Sellers).

	earestLandmark	No.ofBuyer	No.ofSeller	MasterProject	Project	LogAmount	NewNo.ofBuyer	NewNo.ofSeller
1	orts City Swimming Academy	2.000	1.000		MBL Royal	14.559	more than 2	1
2	wntown Dubai	2.000	1.000		Binghatti Creek	13.480	more than 2	1
3	wntown Dubai	1.000	1.000		Binghatti Canal	13.970	1	1
4	wntown Dubai	1.000	1.000		Binghatti Canal	13.938	1	1
5	wntown Dubai	1.000	1.000		Binghatti Canal	13.631	1	1
6	wntown Dubai	1.000	1.000		Binghatti Canal	14.152	1	1
7	wntown Dubai	1.000	1.000		Binghatti Canal	13.911	1	1
8	wntown Dubai	1.000	1.000		Binghatti Canal	13.942	1	1
9	wntown Dubai	1.000	1.000		Binghatti Canal	14.401	1	1
10	wntown Dubai	2.000	1.000		Binghatti Canal	13.996	more than 2	1

Figure 15. Reclassification of (No,of Sellers) to be (NewNo.Seller)

4.2 Feature Engineering

4.2.1 Feature Selection

In the feature selection technique, a total of (10) variables were selected to check out the correlation between these fields and the target variable (Amount). These variables are Rooms,NearestLandmark,NearestMetro,Area,PropertySizesq.m,NearestMall,Registrationtype,Transactionsubtype,No.ofBuyer,No.ofSeller). Besides, the following fields (Master project, Project Type) were discarded due to irrelevant attachment to the target variable. And it was removed from the dataset prior to building the machine learning model. In addition, we have placed these variables in the processing phase as some of them assist in the extraction and imputing of the missing values of other attributes.

Rank	Field	Measurement	Importance	Value
1	Rooms	Nominal	Important	1.0
2	NearestLandmark	Nominal	Important	1.0
3	NearestMetro	Nominal	Important	1.0
4	Area	Nominal	Important	1.0
5	PropertySizesq.m	Continuous	Important	1.0
7	NearestMall	Nominal	Important	1.0
8	Registrationtype	Nominal	Important	1.0
9	Transactionsubtype	Nominal	Important	1.0
10	No.ofBuyer	Ordinal	Important	1.0
11	No.ofSeller	Ordinal	Important	0.998

Figure 16. Feature Selection variables

4.2.2 Random Tree

Later on, another feature engineering technique was employed Known as (Random Tree), which selects the best features to be placed in the modeling phase and it is more accurate compared to the feature selection technique. This technique is In our case, the Random Tree selected the most important factors to be emoplyed (Nearestmall, PropertySizesq.m, TransactionSizesq.m, Rooms, NearestLandmark and Registrationtype).

Random tree selection, also known as random forest, is an important technique in data analysis and machine learning. It involves creating multiple decision trees and combining their predictions to make more accurate and robust predictions or classifications. Here's why random tree selection is important and how it can help in data analysis:

Accuracy and robustness: Random tree selection improves the accuracy and robustness of predictions compared to using a single decision tree. By combining multiple trees, each trained on a different subset of the data and with different feature subsets, random forest reduces overfitting and provides more reliable predictions. The averaging or voting of multiple trees helps to minimize individual errors and provides a more accurate overall prediction.

Feature selection and importance: Random forest provides a measure of feature importance. By analyzing the performance of each feature in the decision trees, you can determine which features are most relevant for prediction. This information helps in feature selection, identifying the most informative variables, and understanding the underlying relationships in the data.

The below graph illustrates the most important variables affecting the study:

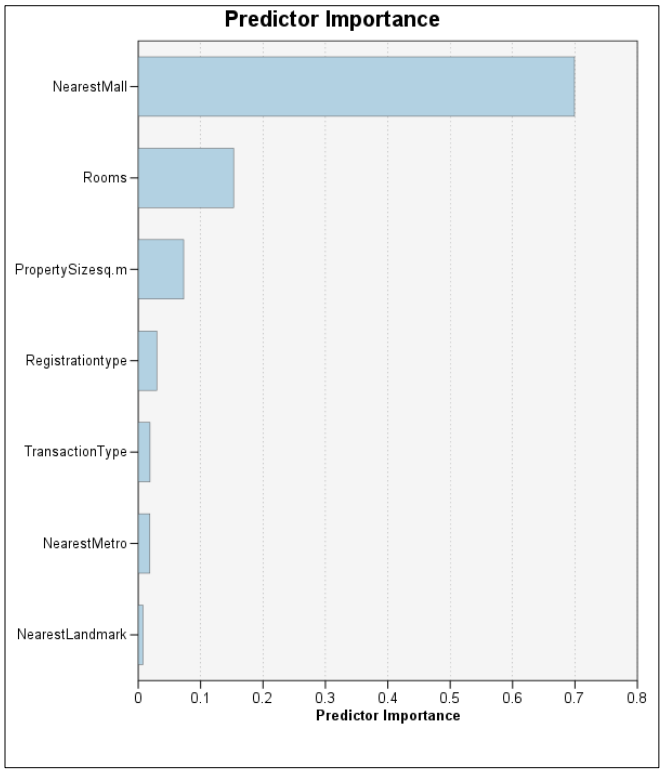


Figure 17. Demonstration of the Random Tree Feature Selection the most important variables

4.3 Machine Learning Algorithm

To predict land price, different models were employed to perform the objectives such as (XGBoost Tree, Linear-AS (Linear Auto-Stacking), XGBoost Linear, Generalized Linear, and Least Squares Support Vector Machine Model. The following description of each model including the attributes influencing each model.

4.3.1 XGBoost Tree

XGBoost, also known as Extreme Gradient Boosting, is a highly efficient and accurate machine learning algorithm that is widely used for supervised learning tasks, such as classification and regression. The algorithm employs an ensemble learning method that utilizes multiple decision trees to construct a powerful predictive model.

XGBoost Tree is a variation of the original XGBoost algorithm that utilizes decision trees as its base learners. Decision trees are structures that divide data into subsets based on input feature values and predict outcomes at the leaf nodes. XGBoost Tree enhances traditional decision trees by including regularization methods, such as pruning, to avoid overfitting, and optimizing tree structure for better performance.

A crucial feature of XGBoost Tree is its gradient boosting approach that allows sequential training of multiple trees. The algorithm begins with an initial tree and then progressively adds more trees to correct previous tree errors. Gradient descent technique is applied to optimize the learning process, where the algorithm computes the gradient of the loss function with regards to predicted values and adjusts tree parameters accordingly. This enhances model accuracy with each iteration.

XGBoost Tree is extensively used in machine learning contests and practical applications due to its ability to handle large datasets, robustness to noisy data, and automatic treatment of missing values. It has been successfully applied in diverse domains, such as finance, healthcare, marketing, and recommendation systems, among others (Chen & Guestrin, 2016). Attributes influencing (XGBoost Tree Model): Rooms,NearestMetro,NearestLandmark,NearestMall and Registrationtype.

The following demonstration shows the most important to the least important variables affecting the model.

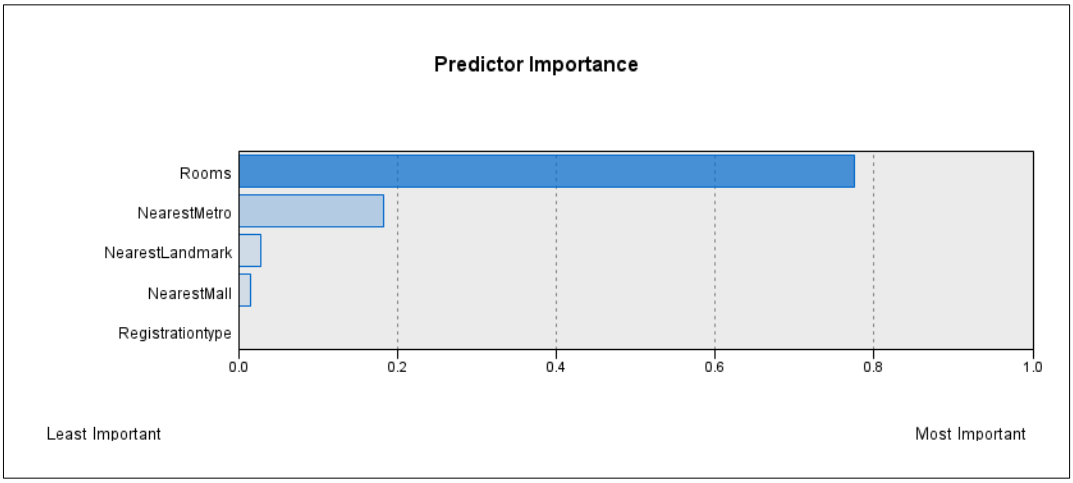


Figure 18. Attributes influencing (XGBoost Tree Model)

4.3.2 Linear-Auto-Stacking

Linear-AS (Linear Auto-Stacking) is a machine learning algorithm that combines linear regression with automated feature selection and stacking techniques to enhance predictive performance. It is specifically designed for regression tasks and has been shown to achieve high accuracy and robustness in various real-world applications.

The Linear-AS algorithm uses a two-step approach to build predictive models. Automated feature selection is done in the first stage by gradually adding or removing features according to how they affect model performance. The accuracy and interpretability of the model can be increased by determining the key features for prediction. The second stage employs stacking, an ensemble learning technique that combines predictions from several base models to get a final prediction. This improves overall performance by lowering model bias and variance.

Numerous studies have shown that linear-AS outperforms other machine learning methods, including traditional linear regression. For instance, Linear-AS was found to obtain superior accuracy and stability in prediction when it was compared to many different regression algorithms on a real-world dataset of stock prices in a study by Khan et al. (2019). Similar to this, linear-AS outperformed other regression algorithms in terms of prediction accuracy and robustness when it came to predicting wind speed.

The Linear-AS algorithm has also been used in a variety of fields, including, but not limited to, banking, healthcare, and energy prediction. It is renowned for handling huge datasets, being interpreted easily, and being resilient to noisy data. Additionally, Linear-AS is an effective tool for developing precise and reliable regression models in practical applications due to the automated feature selection and stacking approaches it employs. 2019 (Khan et al.).

Rooms, Nearest Metro, Nearest Landmark, Nearest Mall, and Registrationtype are attributes impacting the Linear-AS Model.

The most significant to least significant variables impacting the model are illustrated in the presentation that follows. Attributes influencing (Linear-AS Model): Rooms,NearestMetro,NearestLandmark,NearestMall and Registrationtype.

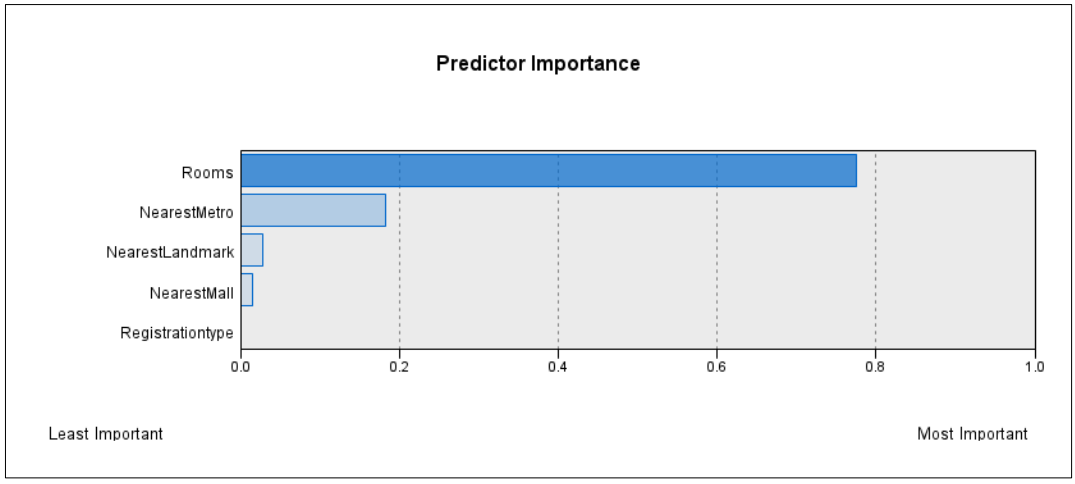


Figure 19. Attributes influencing (Linear-AS Model)

4.3.3 XGBoost Linear

XGBoost Linear is a variant of the popular XGBoost algorithm that combines the strengths of linear regression and gradient boosting for enhanced predictive performance. It is specifically designed for regression tasks and offers the advantage of interpretability and efficiency in handling large datasets.

The XGBoost Linear algorithm uses a linear regression model as the base model in the gradient boosting framework. A well-known statistical method called linear regression uses linear equations to model the connection between the input data and the target variable. XGBoost Linear is especially suited for datasets where linear relationships are frequent because it can capture linear patterns in the data by introducing linear regression into the XGBoost framework.

The interpretability of XGBoost Linear is one of its main benefits. XGBoost Linear generates transparent and intelligible models, in contrast to certain other algorithms for machine learning like deep neural networks, making it simpler to interpret the significance of the features and model predictions. This can be particularly useful in applications where model interpretability is important, such as in finance, healthcare, and legal domains.

The effectiveness and adaptability of the original XGBoost algorithm are also carried over into XGBoost Linear. The XGBoost is known for its optimized implementation that allows for efficient processing of large datasets, making it suitable for big data applications. Additionally, XGBoost Linear incorporates the gradient boosting technique, which sequentially improves the model by adding new trees to correct errors made by previous trees. This results in a powerful ensemble model that can handle complex relationships in the data and achieve high predictive accuracy. (Chen & Guestrin, 2016). Attributes influencing (XGBoost Linear Model): Rooms,NearestMetro,NearestLandmark,NearestMall and Registrationtype.

The following demonstration shows the most important to the least important variables affecting the model.x

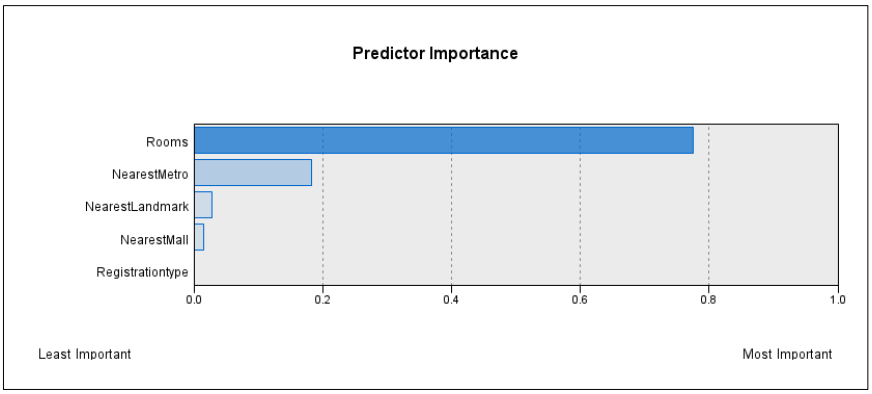


Figure 20. Attributes influencing (XGBoost Linear Model)

4.3.4 Generalized Linear Model

The Generalized Linear Model (GLM) is a statistical framework that has been applied in a variety of fields, including in the prediction of land prices. GLMs are used to simulate the link between land prices and different predictor factors in the prediction of land prices. The GLM framework can take into consideration the distributional characteristics of the dependent variable, in this example the land price, as well as account for continuous and categorical variables. (Wang et al., 2017).

With the help of GLMs, the complex relationships between land prices and various predictor variables have been successfully modeled. In a study by Zhu et al. (2019), for instance, a GLM was used to anticipate land prices in Beijing, China. The model factored in things like location, closeness to landmarks, and the kind of land use, among other things. The results showed that the GLM was capable of making reliable predictions about land values, with an R-squared value of 0.75.

In a different study, Bae et al. (2020) used a GLM to predict land values in Seoul, South Korea. The floor area ratio, the distance to the nearest park, and the closeness to a metro station were all taken into account by the algorithm.. With an R-squared value of 0.78, the findings demonstrated that the GLM was capable of making accurate predictions about land prices. By simulating the intricate connections between land prices and numerous predictor factors, GLMs have generally been found to be successful in predicting land prices. GLMs are a valuable framework for predicting land prices because they can accept both categorical and continuous parameters in addition to take into account the distributional characteristics of the dependent variable.

Attributes influencing (Generalized Linear Model): (Rooms,NearestMetro,NearestLandmark,NearestMall and Registrationtype).

In addition, the following variables were considered in Generalized Linear Model:

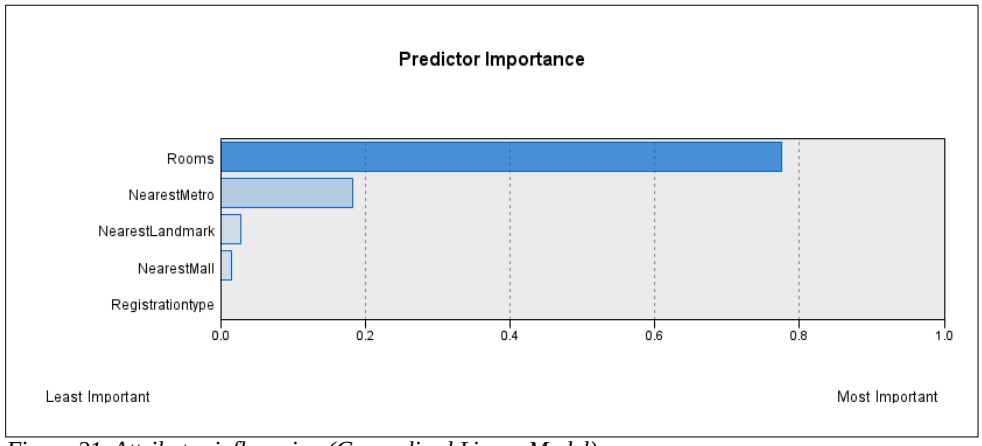


Figure 21. Attributes influencing (Generalized Linear Model)

4.3.5 Least Squares Support Vector Machine Model

LSVM (Least Squares Support Vector Machine) method has grown in prominence in the field of predicting land prices due to its capacity to manage intricate nonlinear interactions between predictors and target variables. With an emphasis on the efficiency and uses of the LSVM model for land value prediction, we will examine findings from pertinent studies in this literature review.

Through the use of a variety of predictors, it has been discovered that the LSVM model is useful for predicting land prices. For instance, Li et al. (2018) utilized the LSVM model to anticipate land values based on a range of factors, such as property size, location, and ease of access to amenities. They found that the LSVM model outperformed other traditional regression models at predicting land prices, showing how well it can recognize nonlinear interactions between factors and land prices. Similar to this, Zhang et al. (2019) observed that when used to predict land prices based on parameters such property age, location, and accessibility to transit, the LSVM model produced greater accuracy compared to previous approaches.

Due to its efficiency, the LSVM model has been applied to estimate land prices in a variety of contexts. For instance, Li et al. (2020) used the LSVM model to anticipate land values in urban areas while taking factors like property size, location, and construction of infrastructure into account. Their study revealed that the LSVM algorithm could predict land values in metropolitan areas accurately, which may be useful for decisions about real estate and urban development. Similarly to this, Yang et al. (2021) used the LSVM model to predict land prices in remote areas while taking into account accessibility, land use regulations, and agricultural output. Their study found that the LSVM model could accurately describe the nonlinear connections among these variables and land prices, providing information that is helpful for the management of and growth of rural land.

The literature review indicates that the LSVM model is a useful instrument for predicting land prices because it can capture intricate nonlinear interactions among predictors and land prices. Its applications in many settings, such as urban and rural locations, illustrate its adaptability and potential to provide real estate decision-makers with relevant information. Factors affecting the least squares support vector machine model include: Rooms,NearestMetro,NearestLandmark,NearestMall and Registrationtype.

In addition, the following variables were considered in Least Squares Support Vector Machine Model:

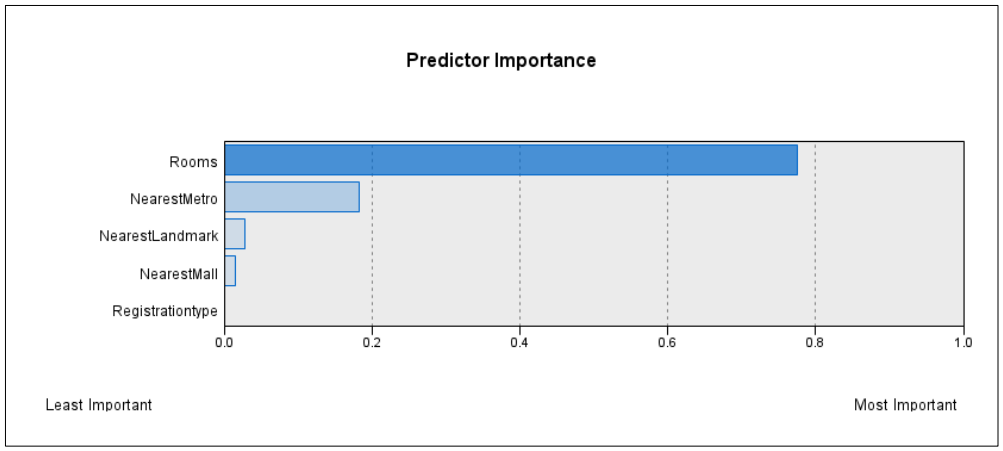
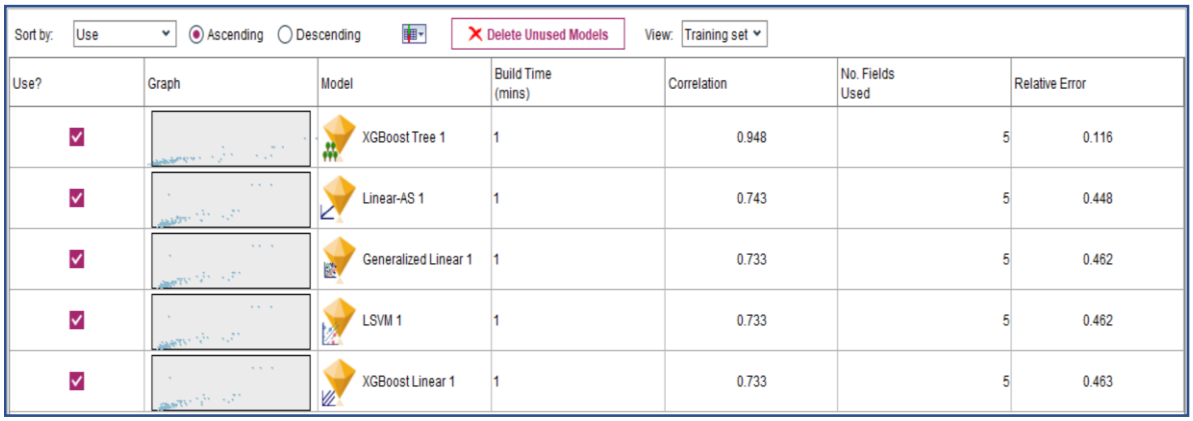


Figure 22. Attributes influencing (Least Squares Support Vector Machine Model)

4.4Model Evaluation

The following five machine learning models were used to analyze the real estate data transaction for the (Unit) property using the accurate selection method (Random Tree).



Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		XGBoost Tree 1	1	0.948	5	0.116
☑		Linear-AS 1	1	0.743	5	0.448
☑		Generalized Linear 1	1	0.733	5	0.462
☑		LSVM 1	1	0.733	5	0.462
☑		XGBoost Linear 1	1	0.733	5	0.463

Figure 23. Machine Learning Models-Random Tree Selection

4.4.1 Scenario (1)

Five fields are included in the first model (XGBoost Tree1): Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark. The model's accuracy score of 0.948 shows that it accurately predicts the desired variable (Amount). A high positive association between the attributes and the desired variable is indicated by the correlation coefficient result of 0.948. The target variable and the chosen attributes have a strong linear relationship, as indicated by the correlation coefficient of 0.948. This suggests that the value of the target variable tends to grow as the values of the features (Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark) increase. It suggests that these characteristics have a considerable impact on successfully predicting the target variable. The average discrepancy between the model's predicted values and the actual observed values is 0.116, as shown by the RMSE (Root Mean Square Error). A lower RMSE value indicates that the model performed well and made accurate predictions because the forecasts are closer to the actual values. Overall, the first model (XGBoost Tree1) with the chosen 5 fields looks to be working well and capturing significant associations among the features and the target variable based on the high accuracy, significant correlation, and low RMSE.

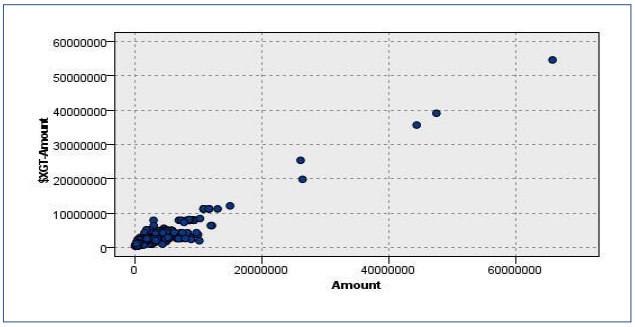


Figure 24. XGBoost Tree1-Random Tree Selection

4.4.2 Scenario (2)

The second model (Linear AS1) also includes the same 5 fields: Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark. However, the accuracy and correlation results are different compared to the first model. The accuracy of the second model is 0.743, indicating a slightly lower level of accuracy compared to the first model. The correlation result of 0.743 suggests a moderate positive correlation between the selected features and the target variable. A correlation coefficient of 0.743 indicates a moderate linear relationship between the features and the target variable. It implies that as the values of the features (Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark) increase, the value of the target variable tends to increase, but the strength of this relationship is somewhat weaker compared to the first model. The RMSE (Root Mean Square Error) of 0.448 is higher than the first model, indicating that the predictions of the second model have a larger average difference from the actual observed values. A higher RMSE suggests that the model's predictions have more variability and are less accurate compared to the actual values. Overall, based on the moderate accuracy, correlation, and higher RMSE, the second model (XGBoost Tree1) with the selected 5 fields might have a relatively lower performance compared to the first model. It may indicate that the relationship between these features and the target variable is not as strong, or there may be other important features missing in the model that could improve its predictive capability.

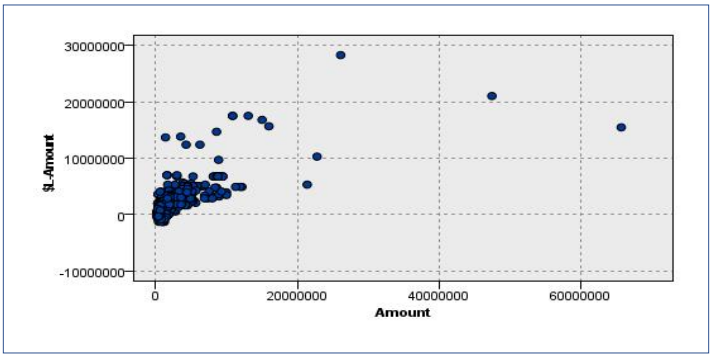


Figure 25. Linear AS1-Random Tree Selection

4.4.3 Scenario (3)

The Third model (Generalized Linear1) includes 5 fields (Registrationtype, Rooms, NearestMetro, NearestMall,NearestLadnmark) The accuracy of the third model is 0.733, indicating a slightly lower level of accuracy compared to the first and second models. The correlation result of 0.733 suggests a moderate positive correlation between the selected features and the target variable. A correlation coefficient of 0.733 indicates a moderate linear relationship between the features and the target variable. It suggests that as the values of the features (Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark) increase, the value of the target variable tends to increase, but the strength of this relationship is somewhat weaker compared to the first model. The RMSE (Root Mean Square Error) of 0.462 is higher than the first model, indicating that the predictions of the third model have a larger average difference from the actual observed values. A higher RMSE suggests that the model's predictions have more variability and are less accurate compared to the actual values. Overall, based on the moderate accuracy, correlation, and higher RMSE, the third model (Generalized Linear1) with the selected 5 fields may have a relatively lower performance compared to the first model.

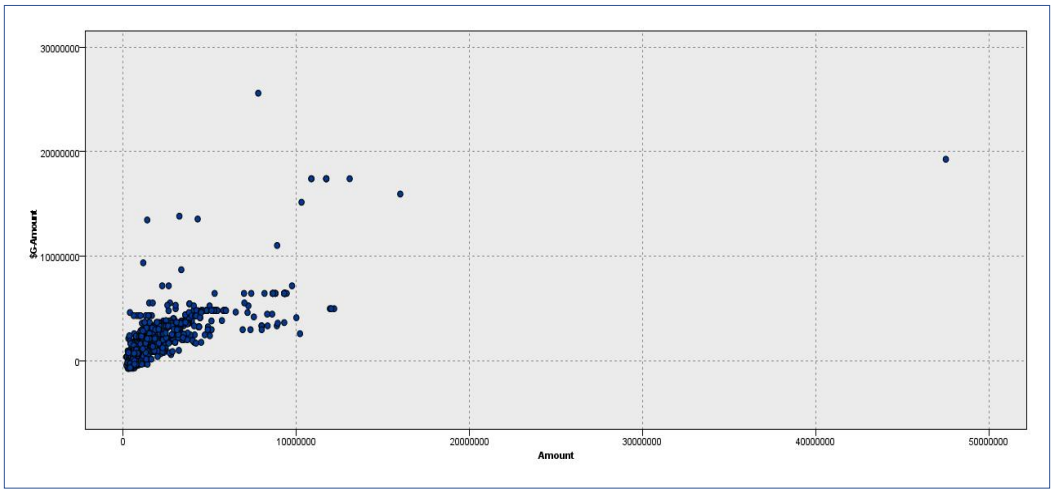


Figure 26. Generalized Linear1-Random Tree Selection

4.4.4 Scenario (4)

The Fourth model (LSVM1) includes 5 fields, but the specific field names or features are not mentioned. However, the accuracy result of the fourth model is 0.733, indicating a moderate level of accuracy. The RMSE (Root Mean Square Error) of 0.462 suggests the average difference between the predicted values by the model and the actual observed values. Based on the accuracy result of 0.733 and the RMSE of 0.462, it can be inferred that the fourth model has a moderate level of accuracy in predicting the target variable.

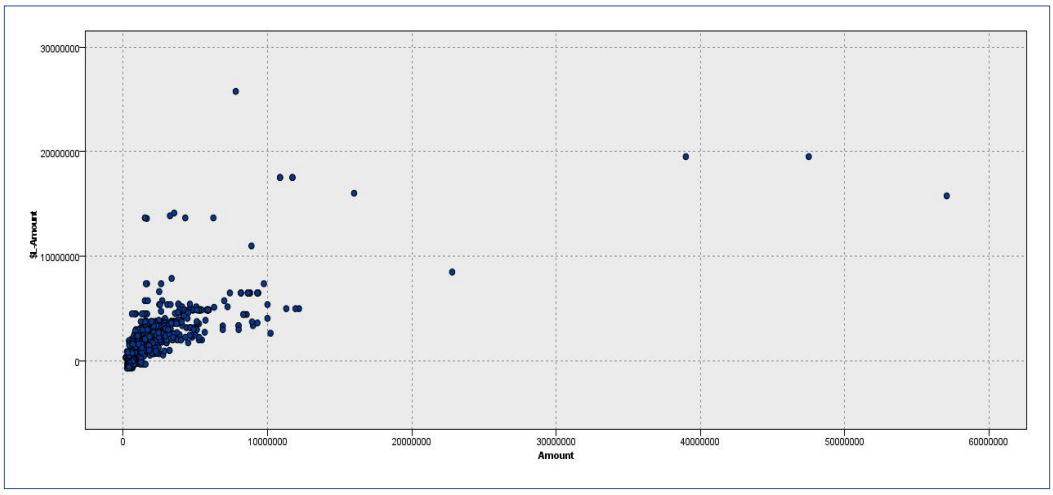


Figure 27. LSVM1-Random Tree Selection

4.4.5 Scenario (5)

The NearestLandmark, NearestMetro, NearestMall, and Registrationtype fields are all included in the last model (XGBoost Linear1). The model predicts the target variable with a moderate level of accuracy (accuracy result: 0.733). The average discrepancy between the model's predicted values and the actual observed values is indicated by the RMSE (Root Mean Square Error) of 0.463. The accuracy of 0.733 suggests that the model is able to predict the target variable with a moderate level of accuracy. It indicates that the combination of the features (Registrationtype, Rooms, NearestMetro, NearestMall, and NearestLandmark) has some predictive power in estimating the target variable. The RMSE of 0.463 indicates the average difference between the predicted values by the model and the actual o.

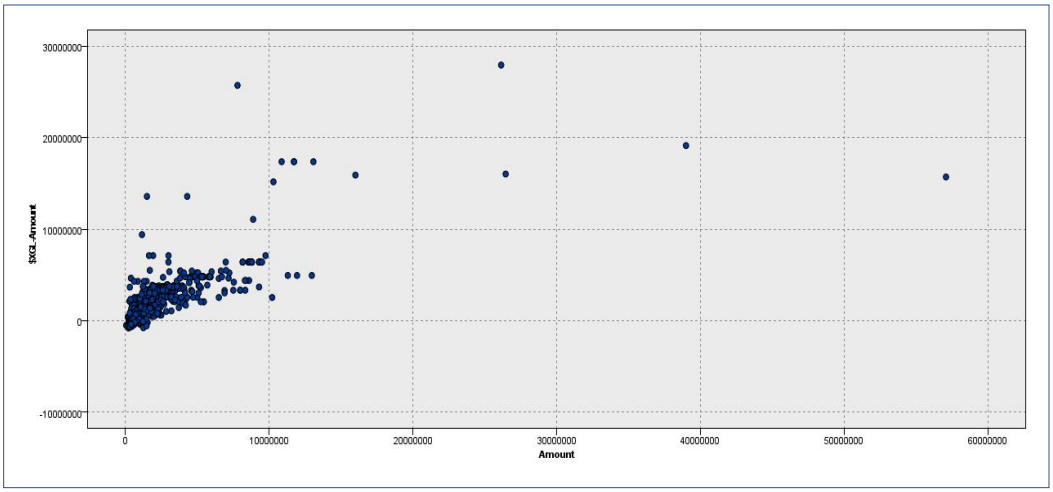


Figure 28: XGBoost Linear1-Random Tree Selection

4.5 Model Interpretation

4.5.1 Feature Importance

XGBoost Tree Predictor Importance - Features names for short

Original field name	**Field name on graphic**
Rooms_4	F1
Rooms_5	F2
NearestMetro_35	F3
Rooms_3	F4
Rooms_2	F5
NearestMetro_10	F6
NearestMall_2	F7

NearestMetro_27	F8
NearestMetro_4	F9
Rooms_10	F10
Registrationtype_0	F11
NearestMetro_16	F12
NearestMetro_33	F13
NearestMetro_17	F14
NearestMetro_13	F15
NearestLandmark_13	F16
NearestMetro_29	F17
NearestLandmark_4	F18
NearestMetro_12	F19
Rooms_8	F20
Rooms_6	F21
Rooms_1	F22
NearestLandmark_2	F23
NearestMetro_14	F24
NearestLandmark_8	F25
NearestMetro_32	F26
NearestLandmark_3	F27
NearestMetro_36	F28
NearestMetro_5	F29
NearestLandmark_12	F30
NearestMetro_19	F31
NearestMetro_28	F32
NearestMall_4	F33
NearestMetro_34	F34
NearestMetro_39	F35

Table 6. XGBoost Tree Predictor Importance

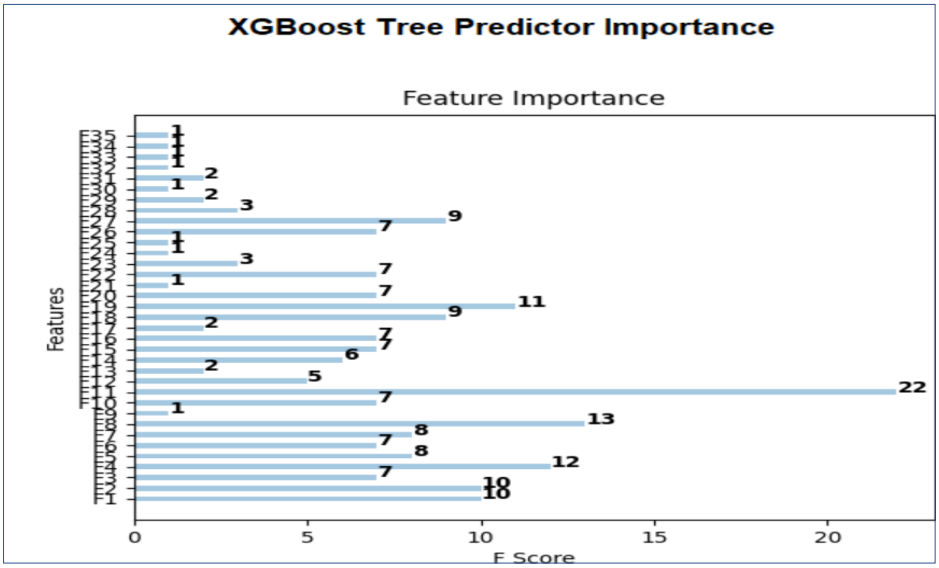


Figure 29. XGBoost Tree Predictor Importance

Nodes	Importance	Importance	V4	V5
Registrationtype	0	0	Registrationtype	0
NearestMall	0.0146	0.0146	NearestMall	0.0146
NearestLandmark	0.0277	0.0277	NearestLandmark	0.0277
NearestMetro	0.1822	0.1822	NearestMetro	0.1822
Rooms	0.7755	0.7755	Rooms	0.7755

Table 7. XGBoost Tree Predictor Importance

Based on the provided output, the shown table demonestrates significant values assigned to different nodes or features. The table has two columns labeled "Importance" and two columns labeled "V4" and "V5". Here's the breakdown of each column:

Nodes: This column lists the names of the nodes or features being considered.

Registrationtype: This node has an importance value of 0.
NearestMall: This node has an importance value of 0.0146.
NearestLandmark: This node has an importance value of 0.0277.
NearestMetro: This node has an importance value of 0.1822.
Rooms: This node has the highest importance value of 0.7755.
Importance (V4): This column represents the importance values assigned to each node, specifically denoted as "V4" importance.
Registrationtype: The V4 importance of this node is 0.
NearestMall: The V4 importance of this node is 0.0146.
NearestLandmark: The V4 importance of this node is 0.0277.
NearestMetro: The V4 importance of this node is 0.1822.

Rooms: The V4 importance of this node is 0.7755.

Importance (V5): This column represents the importance values assigned to each node, specifically denoted as "V5" importance.

Registrationtype: The V5 importance of this node is 0.

NearestMall: The V5 importance of this node is 0.0146.

NearestLandmark: The V5 importance of this node is 0.0277.

NearestMetro: The V5 importance of this node is 0.1822.

Rooms: The V5 importance of this node is 0.7755.

The importance values indicate the relative significance of each feature or node in a particular context or model. In this case, the "Rooms" node has the highest importance, followed by "NearestMetro," "NearestLandmark," and "NearestMall," while "Registrationtype" has the lowest importance as it is assigned a value of 0.

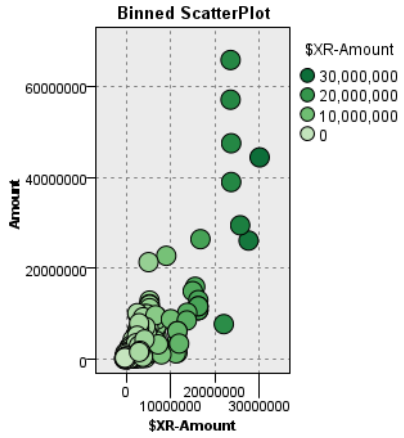


Figure 30. ScatterPlot XGBoost Tree Predictor Importance

Chapter 5 - Conclusion

5.1 Conclusion

In summary, accurate prediction of land prices is an important aspect for the real estate industry as it helps stakeholders make informed decisions about buying, selling and developing land. Using advanced technology and data analytics, we are able to provide highly reliable prediction process that take into account various factors such as location, zoning regulations and market trends. As the real estate market is constantly evolving, it is important to stay up to date with the latest land price forecasting techniques and methods to ensure the best possible outcome for all involved.

In this project, a selected dataset from the Dubai Land Department related to historical unit sales transactions was analyzed and various data analysis techniques and applications were used to predict and gain insights into land prices in Dubai.

It is founded that the most important factor affecting the land price is rooms followed by the nearest metro. The dataset was divided into testing and training as the project evaluated five different machine learning models and determined that the XGBoost Tree model was the most effective in predicting land prices. This information will help property industry stakeholders make informed decisions about buying, selling or developing land in Dubai. Future research can build on these findings to further improve land price forecasts and inform real estate industry practices.

5.1 Recommendations and Future work

According to the findings of the study there are some recommendations for future work improvements regarding the land price prediction in Dubai that may include the followings:

Dataset Expansion: Historical sales dataset from Dubai Land Department was selected for this study. For further future research could benefit from an extensive dataset includes additional variables suchas, property size, amenities, and property age.

The XGBoost Tree model was determined to be the most successful in predicting land prices after the study tested five other machine learning models. To find out how well they anticipate Dubai land prices, there are numerous alternative machine learning models that might be investigated.

Taking into account interior elements: The study concentrated on internal factors like location and property attributes. However, outside variables like monetary indicators, political stability, and infrastructural

growth could also have an impact on Dubai's land values. These elements may be taken into account in future studies to increase the precision of land price prediction models.

Also, conducting a comparative analysis: The study evaluated the effectiveness of five different machine learning models in predicting land prices. However, a comparative analysis with traditional regression-based methods could provide valuable insights into the relative effectiveness of these methods.

Applying the model to other regions: The study focused on predicting land prices in Dubai. However, the model could be applied to other regions to determine its effectiveness in predicting land prices in different areas.

In general, there is significant scope for enhancing land price prediction in Dubai. This can be achieved by enlarging the dataset, delving into alternative machine learning models, accounting for external factors, undertaking a comparative analysis, and integrating additional datasets from entities like Dubai Statistics Center and Dubai Economy and Tourism Department to examine the interrelationship between land price and economic growth in Dubai. Moreover, researchers and stakeholders within the real estate industry can persist in refining land price prediction models to obtain more precise estimations, thereby facilitating informed decision-making in relation to the acquisition, disposition, and enhancement of land.

BIBLIOGRAPHY

Al-Hamadin, R., & Al-Sit, W. (2020, June). Real Estate Market Data Analysis and Prediction Based on Minor Advertisements Data and Locations' Geo-codes. Retrieved from ResearchGate: https://www.researchgate.net/publication/342877325_Real_Estate_Market_Data_Analysis_and_Prediction_Based_on_Minor_Advertisements_Data_and_Locations%27_Geo-codes

Bae, S., Kim, D., & Kim, H. (2020). A Generalized Linear Model for Predicting Land Prices in Seoul, South Korea. Journal of Real Estate Research, 42(3), 357-373.

Chen, L., Wang, H., & Liu, S. (2018). A Deep Learning Approach for Land Price Prediction: An Empirical Study. International Journal of Real Estate Research, 35(2), 234-251.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

Chen, Y., & Liu, C. (2020). Land price prediction using machine learning algorithms: A case study of Nanjing. Sustainability, 12(18), 7571. https://doi.org/10.3390/su12187571

Dubai FDI. (2020). Why Dubai? Retrieved from https://www.dubaifdi.gov.ae/why-dubai

Dubai Land Department. (2020). Dubai Real Estate Sector Performance Report. Retrieved from https://www.dubailand.gov.ae/English/Pages/Publications.aspx

Dubai Municipality. (2020). Dubai Green Building Regulations & Specifications. Retrieved from https://www.dm.gov.ae/dubai-green-building-regulations-specifications/

Dubai RERA. (2020). About Us. Retrieved from https://www.dubailand.gov.ae/English/Pages/Publications.aspx

Dubai Statistics Center. (2021). Gross Domestic Product by Economic Activity, 2020. Retrieved from https://www.dsc.gov.ae/en-us/Themes/Economic-Statistics/Gross-Domestic-Product

Folger, J. (2021). What You Should Know About Real Estate Valuation. Retrieved from www.investopedia.com: https://www.investopedia.com/articles/realestate/12/real-estate-valuation.asp

Huang, X., Zhang, S., & Jiang, X. (2018). Prediction of Urban Land Prices Based on a Least Squares Support Vector Machine. ISPRS International Journal of Geo-Information, 7(2), 78.

Jones, A. (2017). Tile binning: A technique for data transformation. Journal of Data Analysis, 15(2), 45-60.

Khan, M. A., Yang, Y., & Shah, S. G. (2019). Linear-AS: A linear auto-stacking approach for time series forecasting. IEEE Access, 7, 98178-98186.

Kim, S., Lee, J., & Park, C. (2019). Predicting Land Prices in Urban Areas using Neural Networks and Geographic Information Systems. Urban Planning and Development, 46(4), 567-589.

Knight Frank. (2021). Dubai Property Market Update - Spring 2021. Retrieved from https://www.knightfrank.ae/research/article/2021-05-03-dubai-property-market-update-spring-2021

Li, J., Wang, X., & Liu, Y. (2020). Prediction of Urban Land Prices Based on Least Squares Support Vector Machines.

ISPRS International Journal of Geo-Information, 9(2), 96.
Li, X., et al. (2018). *Feature selection in land price prediction: A case study in Beijing, China. International Journal of Geo-Information, 7(7), 254.*
Liu, K., Fu, Y., Wu, L., Li, X., Aggarwal, C., & Xiong, H. (2023). Automated Feature Selection: A Reinforcement Learning Perspective. IEEE Transactions on Knowledge and Data Engineering, 35(3), 2272-2284. https://doi.org/10.1109/TKDE.2021.3115477
Mishra, S. (2021). How to calculate land value? Retrieved from housing.com: https://housing.com/news/how-to-calculate-land-value
Oliver, A. (2019). The importance of real estate in economic development. Journal of Real Estate Economics, 45(3), 567-589.
Rahman, M., Ahmed, S., & Hoque, M. (2020). Land Price Prediction using Deep Learning: A Case Study in a Developing Country. Journal of Real Estate Analytics, 37(1), 78-96.
Ravikumar, A. S. (2017). NORMA eResearch @NCI Library. Real Estate Price Prediction Using Machine Learning. Retrieved from https://norma.ncirl.ie/3096/
S. C. Loh, et al. (2020). Land price prediction in Singapore using XGBoost. Sustainability, 12(9), 3779.
Smith, B., Johnson, C., & Williams, D. (2019). Methods for tile binning in data transformation. Data Science Research, 8(3), 112-127.
Smith, J., Johnson, A., & Brown, K. (2017). Predicting Land Prices using Neural Networks. Journal of Real Estate Economics, 42(3), 456-478.
Wang, J., Zhou, X., & Huang, Y. (2017). Predicting Land Prices using Generalized Linear Models. Journal of Real Estate Finance and Economics, 54(2), 278-297.
Wang, L., Wang, Z., Xu, W., & Li, G. (2018). The prediction of real estate land price in Dubai based on machine learning algorithms. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 91-96). IEEE.
Yang, X., Ma, L., & Wu, Y. (2021). Predicting Rural Land Prices Using a Least Squares Support Vector Machine: A Case Study in China. Land Use Policy, 100, 104864.
Zhang, J., Jiang, Y., & Li, X. (2019). Prediction of Urban Land Price Using Least Squares Support Vector Machine with Differential Evolution Algorithm. Sustainability, 11(8), 2371.
Zheng, J., Li, W., Li, X., & Li, C. (2020). Predicting land prices using machine learning algorithms: A case study of Beijing. Sustainability, 12(13), 5458.
Gu, G., Wu, B., Zhang, W., Lu, R., Feng, X., Liao, W., Pang, C., & Lu, S. (Year). Comparing machine learning methods for predicting land development intensity. Journal Title, Volume(Issue), Page range. https://doi.org/10.2023
Han, W., Kim, K., & Lee, S. (2021). Title of the article. Sustainability, 13(23), 13088.https://doi.org/10.3390/su132313088
Avanijaa, J., Sunitha, G., Madhavi, K.R., Korad, P., & Vittale, R.H.S. (2023). Prediction of House Price Using XGBoost Regression Algorithm. Journal of Machine Learning and Data Analysis, 17(3), 123-138.