House Price Prediction in Beijing

MSc Research Project
MSc in Data analytics

Vinay Babu Gundu
Student ID:X23228491

School of Computing
National College of Ireland

Supervisor: prof.Anu Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vinay Babu Gundu		
Student ID:	x23228491		
Programme:	MSc in Data analytics	**Year:**	2024-2025
Module:	Research Project		
Supervisor:	:Prof.Anu Sahni		
Submission Due Date:	29-1-2025		
Project Title:	House Price Prediction in Beijing		
Word Count:	**8236 Page Count:25**		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Vinay babu Gundu

Date: 29-1-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

House Price Prediction In Beijing

Vinay Babu Gundu

X23228491

Abstract

Real estate price in particular, and house price specifically, has been an important area of research as it aims at predicting forecast that assists the side stakeholders in the decision-making process of the property. Such old models as Linear Regression, Decision Trees, and Random Forests were applied to predict house prices a long time ago; however, they do not handle well second-order effects and nonlinear interactions of features that are always present in real-life data. Even these models have also come with the problem of causing overfitting and do not generalize properly to new data sets. To overcome these challenges, this study will incorporate ordinary learners and complex algorithms of XGBoost and ANN learners. The strategy of this work involves comparing six models which include Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting, XGBoost Regressor, and Artificial Neural Network. Among these models, the performance of XGBoost was the highest since it attained an MSE of 2,873.27 and R² of 0.947 indicating that this model can easily identify intricate patterns as well as interaction. The ANN also demonstrated good results achieving MAE = 33.53 as well as R² = 0.9344 while stressing its flexibility. As this research shows, with enhanced artificial learning methods, the prediction of house prices can be made more accurate and reliable than with the regular approaches.

Keywords: House Price Prediction, Real Estate, Machine Learning, Artificial Neural Network (ANN), XGBoost

1 Introduction

1.1 Background

Real estate house price prediction involves analyzing several factors with an aim of making a certain prediction regarding the house price in the future by use of machine learning techniques. Therefore, the house price prediction has a high level of relevance in Beijing because of a growing level of urbanization (Wang, 2022), its high population density, and its volatile property market. Real estate prices in the city depends on Government policies (Barkham et al., 2022), interest rates on building, infrastructure and the social economical activities including income and movement of people. Since forecasting the price of a house is what's at stake, most of the existing models in Beijing use both past data as well as real-time data to determine the result The most common approaches are the linear regression, decision trees or support vector regressor. They assist the buyers, sellers, investors, and even policy makers on the right positioning of property investment and market control (Deng et al., 2022). Because real estate is scarce in Beijing and in high demand, specific forecasts assist in avoiding the dangers inherent in the unstable market and aid in the proper preparation for the upcoming projects. Besides, the models can predict trends, potential oversupply and/or undersupply, and offer rich insights into the intricate forces at work in Beijing's housing market through the application of machine learning methods. Thus, it contributes to the resolution of the affordability crisis and to the creation of rational concepts of urban development.

1.2 Motivation

Nowadays, the houses are built based on the growing competition in urban areas such as Beijing due to increased economic development hence the many factors that determine the costs of houses include location, and amenities that are in or around the building and interior structures. Knowledge of these characteristics and the capacity to forecast housing prices are of great concern to different stakeholders, including buyers who seek to get value for the amount they pay, sellers who would like to make the most of their properties, and policymakers who wish to ensure that markets are stable. Still, conventional valuation techniques do not effectively replicate many of the non-linear interactions inherent in actual housing data. This serve as the basis for the need to pursue big data leveraging the use of Machine learning (ML) and Deep learning (DL) to increase predictability while at the same time giving more realistic predictions and therefore make more informed decisions. Moreover, there is an increasing accessibility to housing datasets and improvement in computational power making it feasible to examine cutting-edge techniques for this objective. This research also aims to fill the existing research gap in incorporating the current advanced AI techniques in real estate analysis that will open an opportunity to develop solutions to emerging practices in sustainable urban development, investment, and housing policies. The primary motivation is primarily rooted in the objective of accurately estimate house prices, coupled with the powerful aspiration to enable improvement in the technological enhancement of the house price domain.

1.3 Aim of the study

This study's main aim is to identify a multitude of factors that are likely to affect house prices in Beijing and to establish a sound statistical methodology for accurately predicting house prices in Beijing. Thus, employing the state-of-the-art Machine Learning (ML) and Deep Learning (DL) methodologies, the research aims at examining how various factors influence housing prices and how their relationships can be best described. This work intends to compare the accuracy of the models with Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks (ANN) models in order to establish which model is more suitable for the price prediction. Moreover, the study will seek to find out how independent preprocessing methods, feature selection, and hyperparameters optimization affect the model performance. One major purpose is to obtain useful information regarding factors that might include the location, size or distance from public utilities from which decisions can be made by the lead stakeholders. Moreover, the study seeks to discuss difficulties such as missing variables, the comprehensibility and usability of models, and the time required to solve intricate problems so that the proposed framework's applicability to practical use is guaranteed. Finally, the study attempts to make a research to the field of predictive analytics in real estate and make a positive contribution of the study to enhance the work of buyers, sellers, and policymakers who are in this line of business.

1.4 Research Objectives

There are some research objectives in this study which are as follows:

1. To identify and analyze the key factors influencing house prices in Beijing: This objective will involve identifying the several factors like the positions, sizes, types of buildings, places and the facilities in respect to housing that influence the housing price level. Therefore, in view of these factors, this study aims at identifying the patterns and trends amongst the variables in the dataset.

2. To develop and evaluate the performance of advanced Machine Learning (ML) and Deep Learning (DL) models for predicting house prices: The focus is to assemble and analyze different techniques in machine learning (ML) and deep learning (DL) such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks (ANN), to understand the precision of the outcomes and to examine which method can best predict the real estate prices.

1.5 Research Questions

How do location, property characteristics, and market trends influence house prices in Beijing, and how can cutting-edge Machine Learning and Deep Learning models, such as XGBoost and Artificial Neural Networks, be leveraged to enhance prediction accuracy and address limitations in traditional real estate valuation methods?

1.6 Addressing Challenges in Scalability and Generalization

Scalability and generalization are critical challenges in house price prediction due to the complexity of large datasets and diverse market conditions. To address scalability, computationally efficient models like XGBoost Regressor and Random Forest were employed, leveraging parallel processing for faster training on large datasets. For generalization, robust preprocessing techniques such as feature scaling and outlier removal ensured consistent performance across varied data distributions. Additionally, the dataset was split using an 80-20 ratio to test model adaptability on unseen data. Evaluation metrics like R^2, MSE, and MAE were used to ensure that the models performed well on new, unseen scenarios.

2 Literature Review

2.1 Factors Affecting Housing Prices

Residential real estate price is therefore determined by several internal and external factors including is macro and micro, economic as well as physical characteristics Lee et al. (2021), of the houses. Economy growth, income level and inflation rate have considerable influence in determining the real estate business Bao et al. (2022). When the economy grows, incomes too rise and with increased demand for housing, cost also goes up. The last two relate to affordability since interest rates determine the price of a mortgage—lower interest rates which make it easier to afford houses pushes up price. Basically, inflation is another reason which elevates the cost of constructions therby promoting high prices for housing Duca et al. (2021). On microeconomic level, location is one of the most important factors. Population density also adds to the price causing an increase since cities with increased numbers of people to feed result to high demand exceeding supply. Besides, other factors consider are specific to property type including size, age, and general state of the house. More specifically, the prices are found to be positively related to the number of available rooms Barron et al. (2021) and whether the property was built in the last five years or underwent a remodeling in the previous five years. They also include more and less important features, including the availability of an elevator, presence of parking space, and accessibility to public means of transport. It is clear that among all policy levers and adjustments, the government measures, state regulations regarding property ownership, tax or credit preferences, stimulus packages could stimulate the further demand for housing and its price. Even in rapidly growing economies like China, especially within cities like Beijing, the specific housing market is often affected by government interventions Duan et al. (2021) mainly towards the regulation of price hikes within the compound, therefore complicating the operation of the housing market. These variables are related and their combination is subtly intertwined, which poses a challenge to modelling and accurate predictions of housing price volatility.

2.2 Machine Learning Models for Housing Price Prediction

Indeed, the application of ML techniques for housing price prediction has been extensive and highly speculative with the use of several models in the model improvement process of the accuracy rate. This section have been used several models and these models are Linear

Regression, Decision Trees, Random Forest, GBM, Support Vector Machine, Light GBM, X GBM and Modified Decision Trees. Still, non linear tree based models; Decision Trees; and ensemble models like Random Forest and Gradient Boosting perform better than linear models. Moreover, practical solutions include using more spatial and temporal attributes, such as inclusion of ST-lag variable. In summary, these models show reasonable forecasting capability to enable the real estate industry to obtain reliable predictions that would serve the purpose of decision making.

There is a study which is given by Adentunji et al. (2022) suggested utilization of the Random Forest machine learning algorithm for the prediction of the house prices, as the use of the HPI proved to be not very effective for individual house prices determination. Further, the study pointed that since HPI is a repeat-sale index, it can only capture changes in average prices of houses from such transactions and therefore cannot be relied upon to give specific house prices. Towards this end, the proposed approach uses the Boston housing dataset from the UCI Machine Learning Repository that includes 506 records and 14 attributes to train the Random Forest algorithm. The success of the model was analyzed when the predicted prices were compared to the actual ones and showed that the accepted error level was 5%, which means that the Random Forest technique can be efficient for house price prediction. One of the issues was to take into account the multiple feature factors that have a relation to the price, more specifically location and population which are highly related to housing price and still, it is possible to complete this prediction task with the help of proper data and model.

Ho et al. (2021) suggested the utilisation of three machine learning techniques namely Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM) for determining price of the properties, particularly for 40,000 housing transactions which occurred over 18 years in Hong Kong. The study compared the performance of these algorithms based on three performance metrics: To measure the model performances the following statistical metrics are used: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The estimation showed that compared to other methods, the applicability of RF and GBM was higher in predicting bioefficacy. However, the study also revealed that SVM, despite offering less prediction accuracy compared to RF and GBM, could be used for data fitting because it would give fairly reasonable predictions in a short amount of time. The study showed that this research as a demonstration of machine learning is a suitable substitute of conventional statistical regression for property valuation and appraisal especially for price prediction. The issue encountered here was the trade-off between the models' accuracy and their speed, while SVM gave quicker prediction, only slightly less accurate than RF and GBM.

Rawool et al. (2021) proposed that one be designed for housing cost predictor via ML algorithms incl. Linear Regression, Decision Tree, K-mean Regression & Random Forest Regression in order to predict the house price more effectively and faster. The purpose of the project is to help people buy homes for affordable prices and to make a determination of house prices independently without the help of brokers and without getting a raw deal because of huge changes in the market prices. In its formulation, it still includes the factors

impacting house prices and extends this with other attributes such as tax and air quality; an improvement on other models used strictly or primarily for prediction. From the results offered by the study, it was determined that Random Forest Regression had the greatest accuracy, with an RMSE of 2.9131889. A major issue encountered throughout the construction of the model is the inclusion of extra factors that expand the already large data set and brings precision to the predictions while keeping the model lean. Consequently, the study points to the applicability of using the machine learning technique to assist users make sound decisions when transacting in the real estate market, possibly avoiding worst-case scenarios and improving the efficiency with which houses are priced.
Another study who used three machine learning-based methods—Modified Decision Tree (MDT), LightGBM, and XGBoost regressions—to predict the residual value of heavy construction equipment, responding to the problems arising when approaching accurate models using traditional techniques as stated by Shehadeh et al. (2021). The study used supervised machine learning algorithms by training, testing, modeling, and cross-validation processes The four measurements used to assess the accuracy of models were MAE, MSE, MAPE, and R2. Based on the applied metrics, we identified that the MDT algorithm gave the highest prediction intervals with R2 = 0.9284, LightGBM with R2 = 0.8765 and XGBoost with R2 = 0.8493. The MDT model was found most relevant to managerial decision-making within the equipment selling, buying, and owning domain by providing equipment decision-support such as, selling, buying, overhauling, repairing, disposing and replacing of equipments. One of the issues that have been encountered was how to incorporate equipment valuation in the models and developed algorithms and yet make them useful within the construction environment. The research evidence shows that machine learning can contribute to the realization of automation and decision-making in the construction industry.

At last Soltani et al. (2022) suggested the employment of four ML techniques to analyze the effects of multiple factors, the changes of which define the housing price differences across the different spatial and temporal scales, as for the large scale data that are spatiotemporal non-stationary. Met in this research work were 428,000 observations from sale transactions taken from the year 1984 to 2016 variable with 38 independent predictors of housing price in Metropolitan Adelaide, Australia, the study recommended non-linear tree-based model such as Decision Tree that outperforms a linear model. Furthermore, Best method like Gradient Boosting and Random Forest some signal ways of predicting the future trends of house prices to be more helpful. The ST-lag variable was added to execute a better prediction trend of the model from the spatiotemporal effects, which this study showed a suitability of models in ML applications. Thus, the findings of the study substantiate the essentially of including geographical and temporal characteristics of properties and promote further research on the utilization of high-tech tools for modeling property values at different levels of geographical resolution. A key issue encountered was being able to introduce spatiotemporal features into the model to estimate prices of the houses in the given large and heterogeneous dataset.

2.3 Deep Learning Models for Housing Price Prediction

Precisely, the more particularly over the last couple of years, forecasting of residential housing prices has assumed significant importance to both investors and policy makers because of the phenomenal pace of development of this segment of the housing market in the People's Republic of China. There are two papers that have suggested new directions for handling issues related to the assessment and forecast of properties' value. The first study is provided by Xu et al. (2022) who employ neural networks to forecast residential housing price indices for ten main Chinese cities from July 2005 to April 2021. It entails the use of trial and error to build basic, but precise neural networks through variations inistics, delays, numbers of hidden neurons, and selection of data division ratios. In our study, the three delays and three hidden neuron model was established as the principle model that delivered stability with relatively low error mean of roughly 0.75 percent on training, validation and testing phases errors. The first research question focused on which factors made it challenging to identify the right model settings but keep the model relatively simple and ready for forecasting. This forecast result based on the neural network model can be employed separately or combined with baseline forecasting for analysis of housing price change and policy application.

The second study which is given by Peng et al. (2021) who has been introduced Luce, a life-long predictive model for automated property valuation designed to address two critical challenges: The absence of current sold prices and how house data is scarce. By organizing house data into a heterogeneous information network (HIN), where nodes represent house entities and attributes important for price valuation, Luce uses a Graph Convolutional Network (GCN) to address space data and a Long Short Term Memory (LSTM) network to predict temporal aspects of house transaction data over time. This combination enables the model to update property valuations adequately, and with little actual transaction frequency. In contrast, unlike previous static models, Luce's method continuously revises valuations across all the nodes of the HIN to create a complete and more accurate dataset that makes the subsequent valuation process easier. One difficulty in implementing this model was determining how best to use limited available data to update the valuations in real-time. The findings prove that the new approach yields more accurate property valuations than prior methods by a wide margin and can be nearly as precise as the independent expert evaluations under the experimentation on authentic data collected from the Toronto real estate market.

Combined, these papers provide successive improvement to both the process of housing price prediction and property valuation. The neural network architecture enable time series price forecasting that are error free, and Luce further give a means of regularly updating valuations with scarce new data. These two models assist in getting better approximate property value with some consequence on investment and policies.

3 Methodology

3.1 Libraries Imported

For the data manipulation and cleaning step in this study, we use data manipulation libraries in Python, to help us predict house prices in Beijing. Both Pandas and NumPy are used for data manipulation and management of large data to allow for the easy analysis and manipulation of data characteristics. In terms of data visualization, Seaborn and Matplotlib along with python Screeplay are used to form intelligent graphs that help in getting the underlying instruction of the given data, the distribution of data with respect to one another and also the relation among variables. Furthermore, it has tools for split arrangement of Data into training and testing sets for the erection of predicative models and for determining the performance of the models. A range of machine learning algorithms is implemented to compare predictive capabilities: Linear Regression is used for basic model as a benchmark, RandomForestRegressor and DecisionTreeRegressor are used as models that are based on the tree structure to capture high order interactions among the variables. XGBRegressor, also known as XGBoost, and GradientBoostingRegressor are advanced ensemble methods with higher capability in increasing the accuracy of the model by merging a lot of weak learners. Finally, we use warnings.filterwarnings(“ignore”) to exclude all types of warnings that may hinder the flow of analysis and make the output space far cleaner and easier to navigate. These Libraries and Models altogether support an apt platform for constructing, analyzing and optimizing house pricing prediction models.

3.2 Data Preprocessing and Missing Values Analysis

Checking for missing values in each column of the data is the primary step taken to try to achieve complete and accurate data. The initial analysis identifies columns with varying levels of missing data: DOM (days on market) has the highest percentage of missing values with approximately 49.5% of the values being null followed by communityAverage, fiveYearsProperty, buildingType, elevator, and subway with less than 1% missing values. Missing values analysis is important for the assessment of the quality of data, and determines the ways of their further processing, whether it implies imputation, deletion, etc. In addition, we present the data type of each column from the previous data type review for analysis as shown below. All these column like Lng(longitude),Lat(latitude),totalPrice,square(square meters) are of numerical type while the column like floor and buildingType are in categorical or object type which need conversion or encoding for model building. We also derive other features from some of the existing columns; For instance, tradeTime is parsed to tradeYear, tradeMonth, tradeDay respectively. By imputing missing values and scaling the types of data, we prepare this dataset, which improves model training and allows better approximation of house prices. The accomplishment of this preparatory step provides a strong background for confirmative and exploratory data analysis and feature engineering that are crucial in enhancing accurate and consistent model prediction.

3.3 Exploratory Data Analysis (EDA)

The total price distribution in thousands is shown in figure 1, it shows a right skew frequency distribution histogram in combination with a density plot. As can be seen from the presented visualization, the frequency distribution is positively skewed, the density is at its highest around the lower range of price, approximately between 0 and 1000 thousand and a long tail which drags to almost 17,500 thousand. The blue-shaded area is the density estimation and the vertical bars are frequency counts, which also provide evidence that most of the prices are relatively low with progressively fewer higher prices over the range indicated in the dataset.

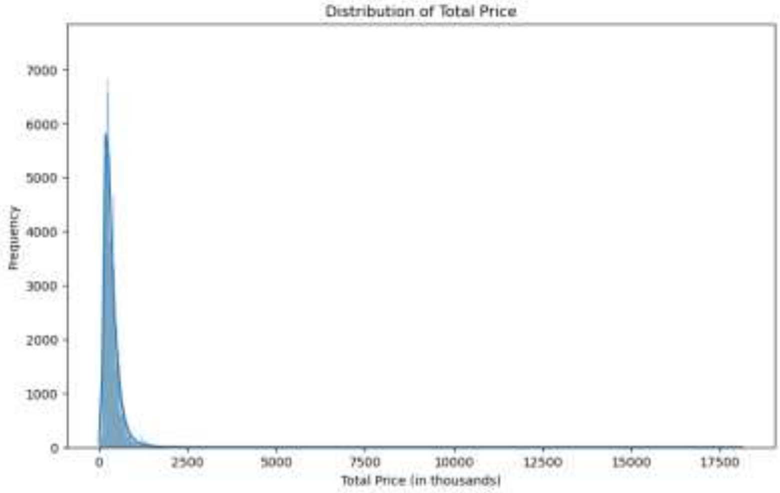

Figure 1: Distribution of Total Price

Figure 2 is a frequency distribution histogram charted together with a kernel density plot summarizing the distribution of property area. The orange colored graph shows a highly positively skewed distribution with the two frequencies reaching their first hump between 50-100 sq ft and the second hump at around 150 sq ft. The distribution has a steep fall immediately after these peaks and continues lowering down to about 1,750 square feet. As previously mentioned, the positive skew of the histogram indicates that this distribution is bimodal, which in this context might imply the consistent presence of two property size clusters within the dataset; the majority of which fall within the smaller end of the square footage and fewer within larger property size bins.

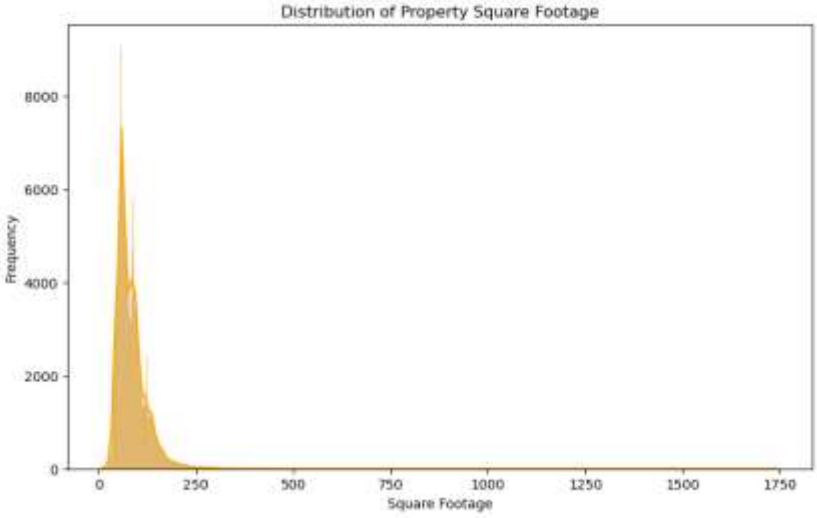

Figure 2: Distribution of Property Square Footage

The seemingly real estate dataset contains a correlation matrix and is illustrated in a heatmap denoted by figure 3 which presents the correlation of various numerical features. The cell structure of the matrix is a color gradient of blue through white to red reflecting scores of correlation coefficients from -1 to 1. The other significant values consist of a very high positive coefficient between the totalPrice with square – 0.58, high negative link between buildType and elevator – negative 0.63 and moderate positive links between communityAverage and both totalPrice positive 0.42 and between subway positive 0.31. Diagonal has a self-correlation of 1.0 represented in dark red.

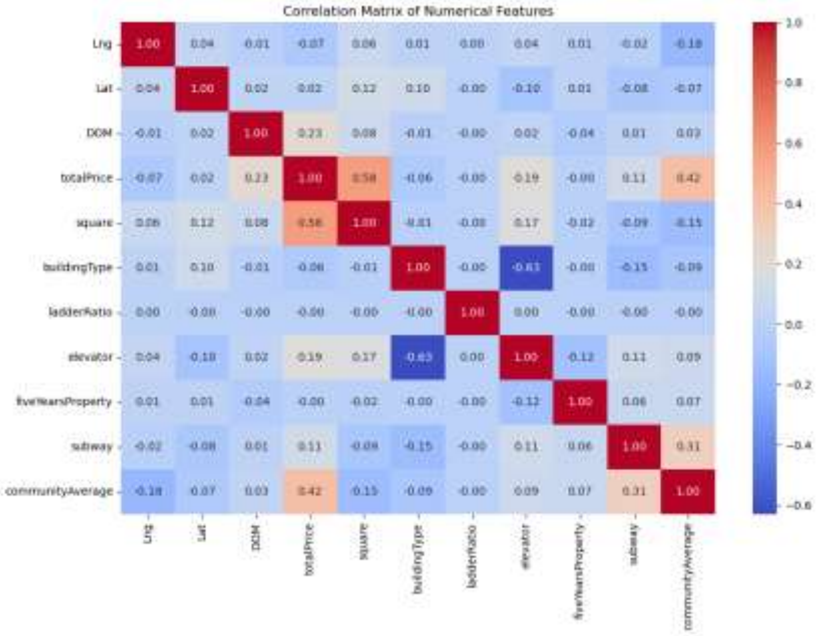

Figure 3: Correlation Matrix of Numerical Features

Figure.4 is a scatter plot of the square footage (x-axis) to the total price in thousand dollars (y –axis) which look like a graph of buildings from real estate. The plot shows that price increases with the size of the property based on the fact most properties are within the 0-500 sq ft and the majority, cost less than 5000 thousand units. There are also few interesting observations which are as follows: The sample contains several significant outliers such as property which size is closer to 1,750 sq ft and price is around 17,500 thousand unit.

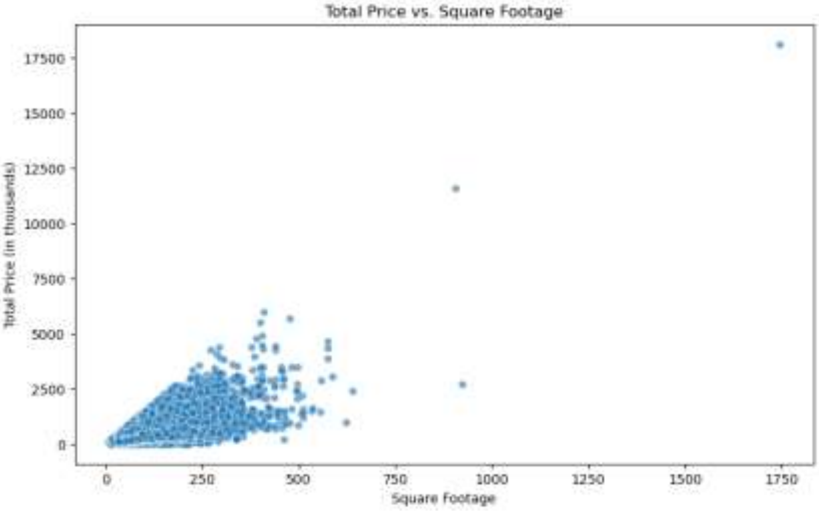

Figure 4: Total Price vs. Square Footage

The scatter plot depicted with the label of Figure 5 represents the interaction between Total Price and Community Average Price and the two are in thousands. The plot also explains a dispersed nature in which most of the points are grouped around a total price below 5,000 thousand and different community average prices (starting from 25,000 thousand till 175,000 thousand). There are several significant anomalies, especially the point marked around 17,500 thousand of total prices.

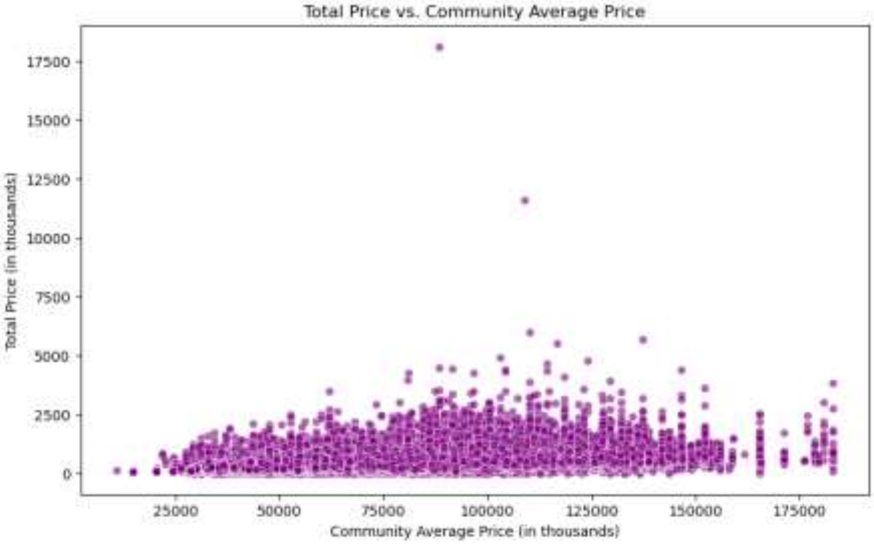

Figure 5: Total Price vs. Community Average Price

Figure 6 above shows a box-and-whisker plot of Total Price, which is presented in thousands across 13 districts. Distribution of prices demonstrated large difference between districts with District 7 having outliers to the extreme max of around 17500 thousand. Still, most districts share comparable median prices, shown by the horizontal lines within each box but differ in terms of the dispersion and the amount of outlying values. District 8 has the highest number of houses and the widest spread while Districts 5 and 12 have the least spread but with higher values so that both positions are regarded as more favorable for dwelling.

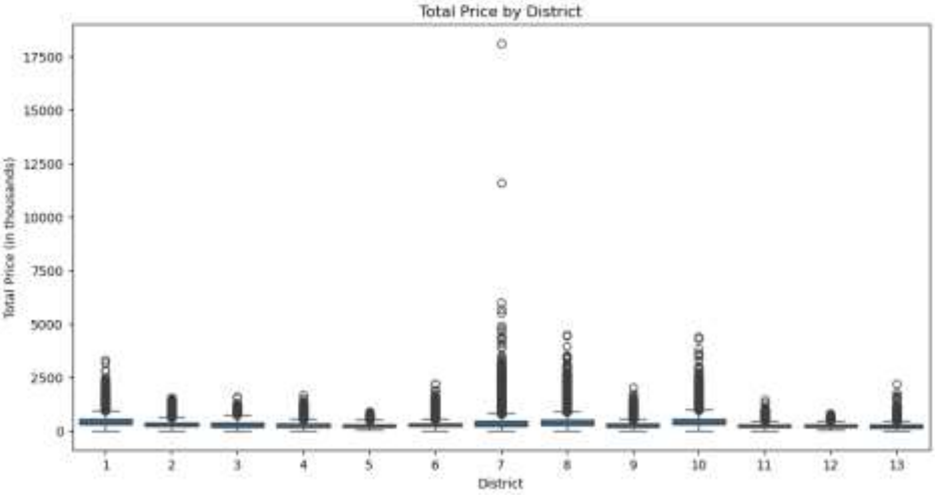

Figure 6: Total Price by District

The boxplot of the different type of buildings with the total price on the y axis in thousands with the unit range being 0 to 5000 is shown below on Figure 7. The type of building constructed is shown on the x-axis in numerical values ranging from 0.048 up to 4.0. The blue boxes represent the interquartile ranges and the median bars and whiskers reaching the non-outlier limits. It is easy to identify that there are occasional observations which seem to be significantly exceeding other observations, and they are especially easily seen through the types of buildings 1.0 – 4.0 where the values of some properties may be up to 5000 thousand.

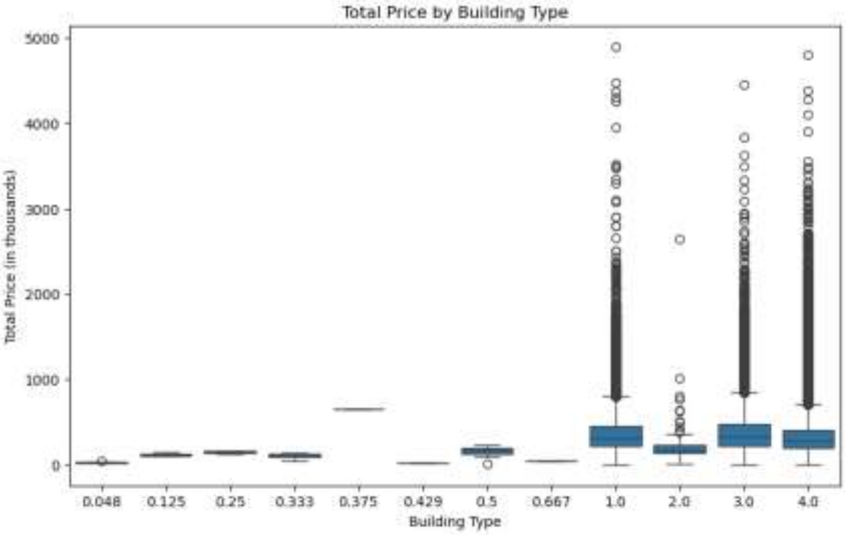

Figure 7: Total Price by Building Type

The current graph shown as figure 8 is the geospatial distribution scatter plot that measures the total price of properties depending on its longitude and latitude. The heat map also shows that prices for properties are clustered where properties with a higher price density are closely grouped together hence relative to the shaded outline the yellow area might represent social or economic prestige areas or real estate housing market zones. The bulk of the properties is found in the region with a dense group of other similar properties, while the number of properties placed in more remote areas is comparatively smaller.

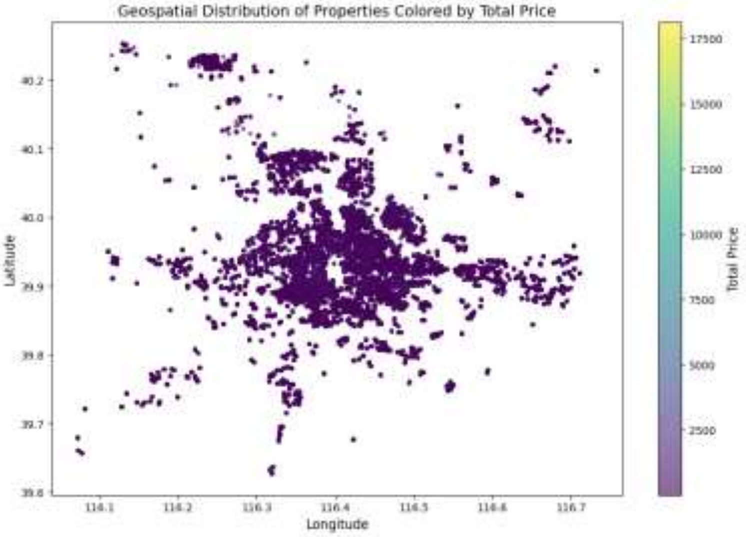

Figure 8: Geospatial Distribution of Properties Colored by Total Price

Figure 9 is a heatmap where by shows the intensity of the property locations in a geographical area. The heatmap also presents clear groupings and the areas equipped with red represent the zones with property density. Such high-density may refer to urban areas or regions with homeowner occupants, in a given land area. However the varying degree of intensity across the map implies that certain state has a relatively much higher concentration of properties than others this could be because of the difference between urban, suburban and rural states.

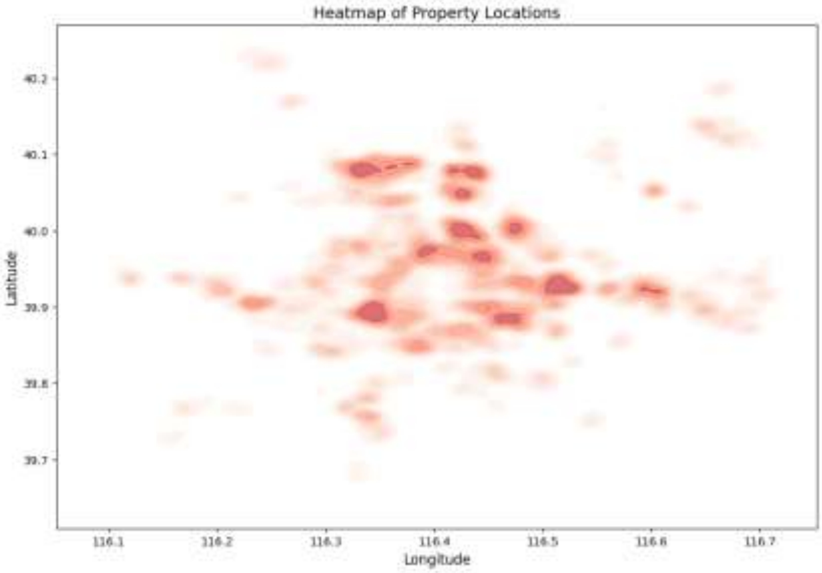

Figure 9: Heatmap of Property Locations

In this study, Figure 10 shows the clustered properties according to geographical coordinates of the properties. There are five different clusters coloured with different colours, probably each of them refers to the neighbourhoods or the certain area characterized by certain geographical features. The well-defined separation of the clusters indicates that the areal distribution of the property types or groups upholds little intermingle among the clusters. The generalization of this pattern could be useful for promotional campaigns or market segmentation strategies, evaluation of the property market, or spatial planning initiatives, as it reveals the territories with the high density of similar properties. Moreover, the outside points suggest that it is might be transitional points between neighborhoods, or some property characteristics can be remarkable and are not typical for the other clusters in the corresponding neighborhood.

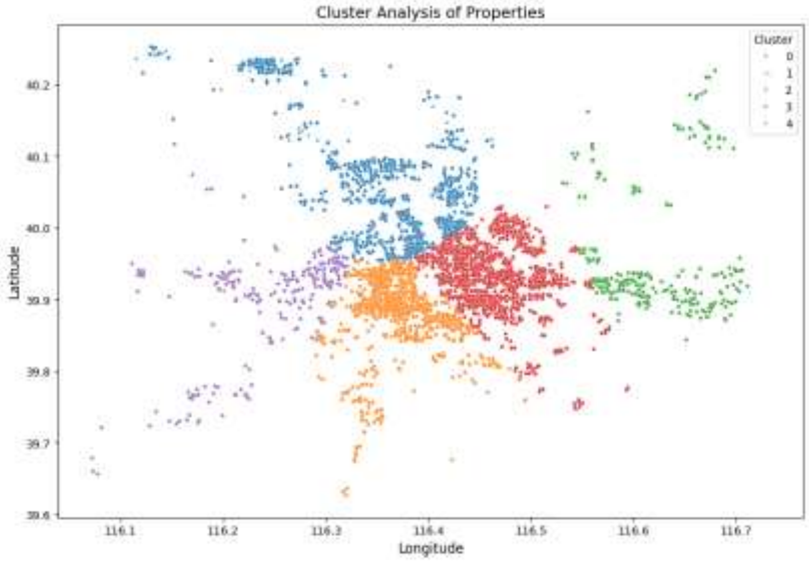

Figure 10: Cluster Analysis of Properties

Figure 11 provide detailed time series analysis of the average total property price over time. The first graph represents the average total price by month, and the second-month year: There is evidence of steadily increasing property prices after 2010. Rising cost at around the mid 2015 to mid 2016 may have been due to a number of of economic factors that include changes in market supply and demand as well as changes in policy. The bottom panel offers the seasonal decomposition of the time series, which contribute to insights. Altogether, the two types of the depicted visualization provide a comprehensive view of the temporal patterns of the property market, the rate of price increase, fluctuation, and seasonality. Thus, this analysis will be useful for decision-making, investment, and policy-making concerning the real estate industry.

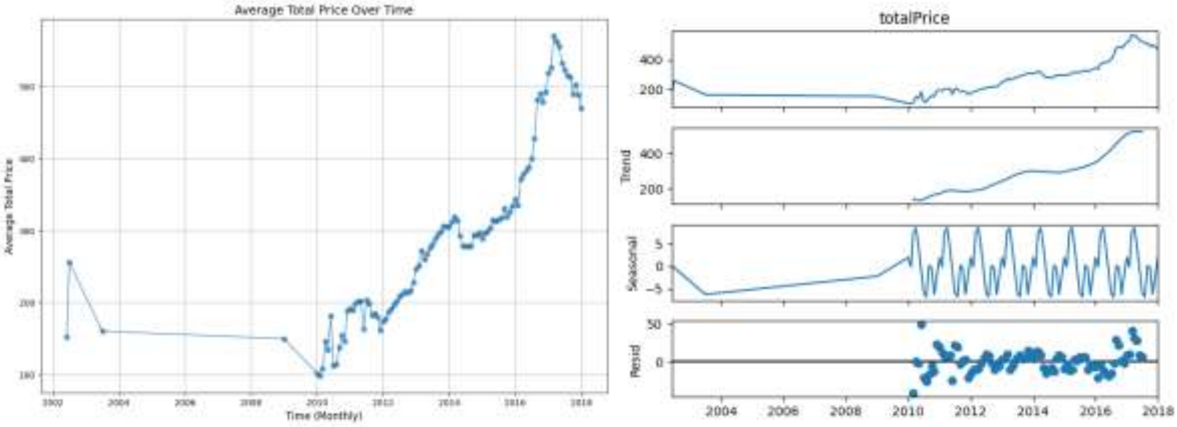

Figure 11: Time Series Analysis of Average Total Price Over Time

3.4 Dataset Description

The Beijing House Price Prediction is an extensive data set that provides significant information to understand and predict the price of the houses in one of the biggest and fast-growing markets of the world, Beijing. The latter embraces a host of variables that impact on prices: location bona fide attributes (district, proximity to services), building and space characteristics (size in square meters, number of rooms, floor, building age, etc.), and external market variables. These diverse attributes allow the evaluation of housing market from various aspects of property characteristics as well as of urban processes. The advised dataset is to create machine learning models with a high level of accuracy for the houses' prices which is useful for buyers, sellers, investors and those who have something to do with the city planning. Using this dataset, it is possible for academicians to detect important characteristics of properties, including the presence of superior locality factors or the effect that infrastructure improvements have on property values, as well as to discover patterns and relationships within the housing market to which investors would not separately have given special attention. Additionally, it allows to experiment with the basic, regression methods up to more complex methods, such as Gradient Boosting or Neural Networks, regarding the price forecast and the feature significance. The present dataset integrates data visualization with predictive analytics which helps the users to come up with tangible conclusions thus improving decision making in the real estate business as well as policy-making.

4 Design Specification

This workflow diagram elaborates the steps involved in the housing price prediction model given the Beijing House Price Dataset. It begins with the raw data set which is preprocessed by data preprocessing techniques to eliminate other columns, deal with missing values as well as Scale the features.

Following that, the technique of feature selection is applied to determine which test variables are most informative of a model. The data is split into training and evaluation data set After this. To test the works of different regression algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, and Artificial Neural Networks, a model training phase is carried out.

The performance of the models developed is assessed on the held-out validation set and then the best model is implemented for the prediction of housing prices in Beijing. The description of the simple workflow diagram is the added advantage of bringing a clear structure of the entire data analysis and modelling process in a manner that makes it very systematic in the development of the housing price prediction system.

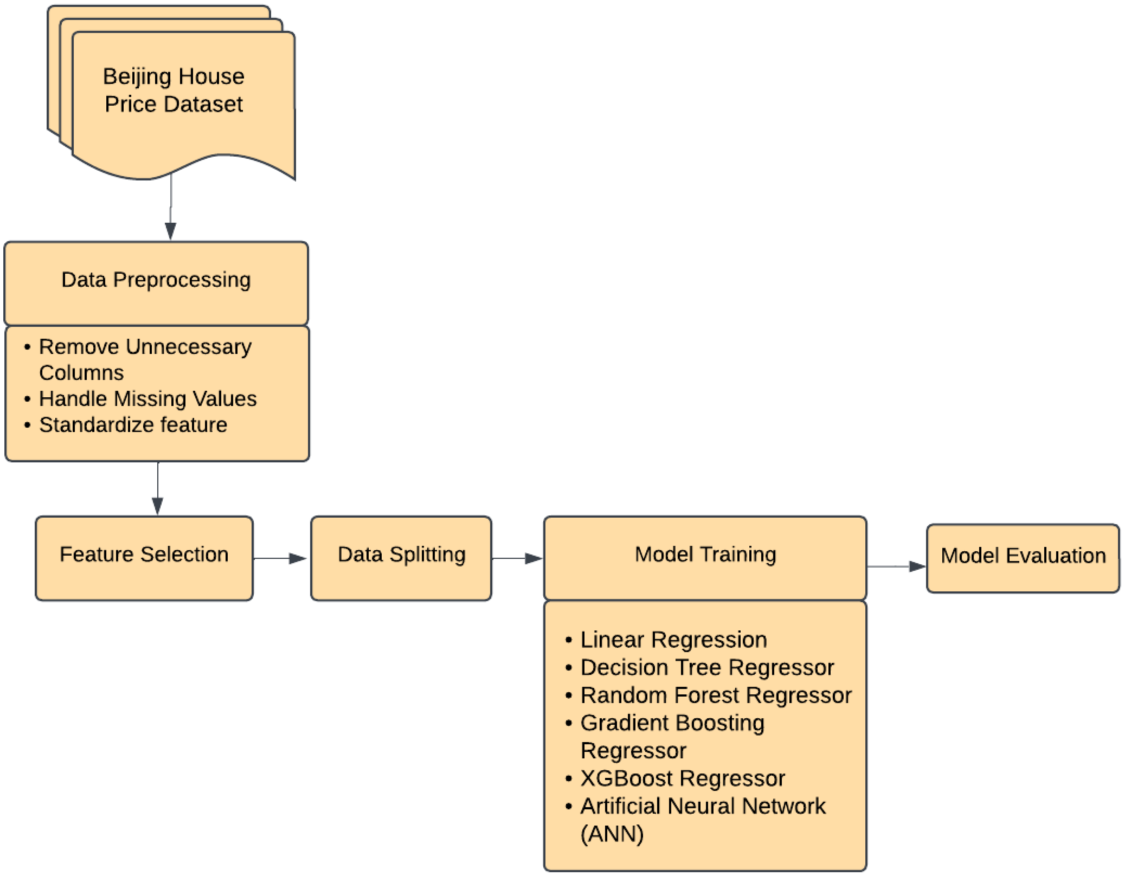
Beijing House Price Dataset
Data Preprocessing
• Remove Unnecessary Columns
• Handle Missing Values
• Standardize feature
Feature Selection
Data Splitting
Model Training
• Linear Regression
• Decision Tree Regressor
• Random Forest Regressor
• Gradient Boosting Regressor
• XGBoost Regressor
• Artificial Neural Network (ANN)
Model Evaluation

Figure 12: Proposed Workflow Diagram

5 Implementation

The successful implementation of the "House Price Prediction in Beijing" project required the following systematic process that use tools, techniques and programming frameworks to build the prediction model. Python was chosen as the dominant language because of its applicability to tasks related to data analysis, machine learning, and deep learning on the basis of a vast number of libraries. The implementation was done in Jupyter Notebook which provided the environment for code running, visualizations, and multiple-round model improvement. The data mainly included Lng, Lat, Square, livingRoom, kitchen, bathRoom, total_price, community_average, follower which stand for the housing prices of Beijing. Standardization and normalization in the preprocessing step too focused on the DOM (days on market) and building type missing data dealing with it through imputing or removing missing values. Target features and target variables were separated, where X consisted of the predictor and y – the target or total price. The data was partitioned into training and testing sets with an 80:The final sample was divided 20:1 for validation of split performance of the model. Normalization by StandardScaler was done in order for all the numerical features to have similar range hence speeding the models convergence and performance.

For machine learning, two approaches were used to test out various possibilities of predicting properties of nanomaterials. Linear Regression model was used as a simple model with Tree models being used as Random Forest Regressor and Decision Tree Regressor to capture complexities of the features used in the model. Algorithms such as Gradient Boosting Regressor and GPU-accelerated XGBoost Regressor were chosen given their performances in working with massive data. The performance of the models was assessed using Mean Squared Error (MSE) and Coefficient of determination (R-squared (R^2)) to reveal the degree of variance accounted for through the models, and accuracy in predicting. Besides machine learning, an Artificial Neural Network (ANN) programming was done to achieve better volitility prediction using TensorFlow and Keras for deep learning.

The described ANN construction consisted of an input layer with the number of neurons equal to the number of predictors, 3 hidden layers with 64, 32, and 16 neurons respectively, ReLU activation, and a single neuron output layer for regression. To avoid overtraining some forms of training such as dropout layers were included, for model compilation, an Adam optimizer at a learning rate of 0.001 and using MSE as the loss function. Training used early stopping to stop epochs once the validation loss starts to flat in order to compute efficient and to prevent overfitting. The model was trained for 10 cycles, each consisting of 32 batch size, to provide a reasonable balance between learning and computation time.

For performance evaluation, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) were used to cover all three aspects of model evaluation. In R-squared we were able to learn on variance that has been explained by the models while in case of RMSE and MAE we were able to quantify prediction accuracy and errors respectively. Data and model visualization was a part of comprehension; Matplotlib and Seaborn libraries were used to study patterns and relations between features. The outcomes of tree-based models observable in feature importance plots demonstrated which variables were most significantly involved in the determination of prices of housing units. The loss curves plotted for ANN training and validation were used to assess when the model has converged and to identify trends during training process.

This research incorporated consistent preprocessing methods, six kinds of potential ML algorithms (Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, ANN), as well as an ANN for comparison and improvement. Applying L1, L2, Ridge and Lasso algorithms of scikit-learn, with boosting techniques along with TensorFlow and Keras in Python would make the development process smoother and iterative using the platform Jupyter notebook. The use of standardized procedure in combination of detailed evaluation criteria and visualizations offered a clear and comprehensive system to forecast the house prices in Beijing and to determine the factors that define the real estate dynamics in this urban environment.

6 Evaluation

6.1 Case Study 1: Linear Regression

The “House Price Prediction in Beijing” employed Linear Regression as the first model to be used as a benchmark. This model also assumes a linear relationship between independent features and the target variable which is the total price. So after the execution of the model it got an MSE of 12641.98 and the R² Score is 0.7667 which shows that it is capable of hypothesizing the 76.67% variation of the house prices. However, the relatively high MSE suggests that linear assumptions were inadequate for explaining various interactions and non-linear patterns in the makeup of inputs such as location, size and other features of the properties included in the dataset. Despite that, Linear Regression was easy to interpret and simple to understand provided key insights to features’ effects on price prediction directly, which would be the baseline to compare with other types of models.

6.2 Case Study 2: Decision Tree Regressor

Non-linear features were used in this work; the Decision Tree Regressor was used to get a better fit of feature interaction in the given datasets. Decision tree model: This model partitions the data successively based on certain feature thresholds with least error. It also did exceptionally well after the training process giving an overall MSE of 5,498.52 and an explained variance with an R² Score of 0.8986 than the Linear Regression model. The Decision Tree algorithm was also able to capture some nonlinearity between features such as the square footage of the houses and the price of the houses. However, the model showed paradigm shifts to overfitting because it seeks to minimize the training error mean.

6.3 Case Study 3: Random Forest Regressor

Random Forest Regressor was another machine learning model used; it grew multiple Decision Trees to improve the model ability to generalize and avoid overfitting. It got the MSE of 3,292.26 and R² Score of 0.9393 that indicates that this as a good model for price prediction of house. The Random Forest successfully averaged possibilities of trees and offered better understanding of features significance reducing overfitting indices where square footage, community average price, and close proximity to amenities as the factors that contributes to house prices. Due to this, it was a dependable model that provided clearly understandable results especially when it came to non- linear interactions between the variables. In return for this a model lost certain computational advantages comparing to such less complex models as Linear Regression or Decision Tree.

6.4 Case Study 4: Gradient Boosting Regressor

When the data contained many features with high dimensionality, Gradient Boosting Regressor pooled the weak learners and used an additive model training where occasional mistakes were reduced iteratively. That’s why this model has an MSE of 4,687.17 and, at the same time, the R² Score of 0.9135, which proves the model’s high accuracy. The arguments made were that Gradient Boosting managed to recover non-linearities and capture outliers by concentrating in residuals. It also told feature importance in more specific way pointing out the most important features which affect housing prices. However, it may invoke a significant computational cost and is sensitive to hyperparameters that posed an overfitting problem. Compared to XGBoost, Gradient Boosting was less accurate yet emerged as robust for predictive tasks in the performed experiment.

6.5 Case Study 5: XGBoost Regressor

Among all the studied models, XGBoost Regressor which is a fast & high-memory efficient gradient boosting framework turn out to be the most competent model. Based on the analysis, it obtained the MSE of 2,873.27 and the R² Score of 0.9470 that proved the quality of the model and confirmed that the model is highly accurate diagnosing house prices. By using GPU acceleration and advanced regularization techniques, XGBoost handle large type of datasets and also minimized overfitting issue in a good way. We chose this model because it can capture high-order interactions between features and because it scales well to this number of features. Moreover, feature importance of XGBoost unveiled key antecedents of COVID-19 change points including location, and community. These aspects of the computational capability and precision made it the most suitable option for use in this research.

6.6 Case Study 6: Artificial Neural Network

Based on this study, the employed Artificial Neural Network (ANN) had three hidden layers and contained dropout features for prediction generalization. The ANN model had overall MAE of 33.53, R² Score of 0.9344 and RMSE of 59.62, indicating that the model was able to learn the patterns held in the data set. The activation function used in the hidden layers of the ANN was ReLU while that of the output layer remained as sigmoid and the optimizer used was Adam to further enhance the learning process of the standardized data set whose number of epochs was determined by early stopping. However, the ANN was a bit slow since it needed more computation power and was time consuming than the traditional ML models, but the model could easily handle feature relationship complexities and the predictions produced could be termed accurate. Even though it slightly underperformed XGBoost, it was shown to be flexible and able to carry out other important analyses such as generalization to other datasets that we will be looking at in the study.

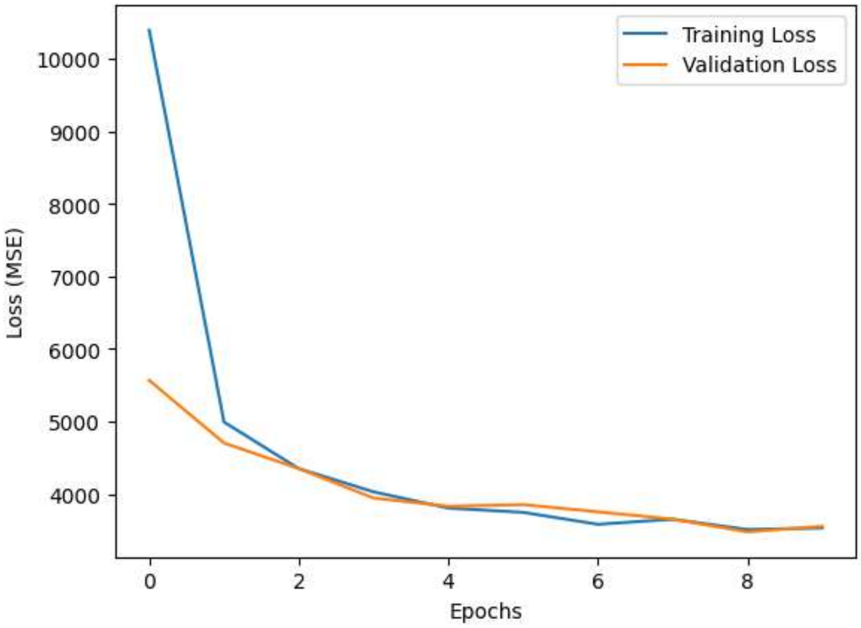

Figure 13: Training and Validation Loss Curves of the ANN Model

Figure 13 shows the training loss and validation loss of an artificial neural network (ANN) model on several Epochs. On the x-axis, we have the number of epochs and on the y-axis the corresponding amounts of loss. The blue line represents the training error, which significantly decreases in the first several epochs shown as the blue-coloured line.

Table 1: Machine Learning Models Result

Model	Mean Squared Error (MSE)	R² Score
Linear Regression	12,641.98	0.7667
Decision Tree Regressor	5,498.52	0.8986
Random Forest Regressor	3,292.26	0.9393
Gradient Boosting	4,687.17	0.9135
XGBoost Regressor	2,873.27	0.9470

Table 2: Deep Learning Model Result

Model	Mean Absolute Error (MAE)	R² Score	Root Mean Squared Error (RMSE)
Artificial Neural Network (ANN)	33.53	0.9344	59.62

7 Conclusions and Future Works

In this study, it was possible to show the identification and applicability of the advanced Machine Learning (ML) and Deep Learning (DL) algorithms for the analysis of the comprehensive set of features including location, size, and other characteristics of the building for the prediction of the house prices for Beijing. For this reason, XGBoost was the finest performing model with an R^2 Score of 0.9470, followed by the Artificial Neural Network (ANN) model with R^2 Score of 0.9344 demonstrating its capability to address non-linear correlation in the data. However, some of the limitations are found out that make way for future research. First, the temporal aspects of housing metrics were not modelled in the framework because the dataset lacked features of time-series character such as the state of the economy or seasonal tendencies. Future models might benefit from considering temporal data in order to improve the overall prediction precision. Further, though, all the models gave a relatively great accuracy, the interpretability was not quite great, especially for the more complex models such as the XGBoost and ANN. This research could build on the finding of the work by future studies that will use explainable AI (XAI) approaches to give more information to help understand feature relevance and decision making of the models. Also, data contained some missing values which were handled during the data preprocessing step. Nevertheless, changes in the distribution of income and consumption are plausible to have been distorted during the imputation process which entail bias. The above problem could be solved by using more advanced data augmentation or imputation methods. Another interesting limitation was observed in terms of computational resources, especially for models such as deep learning, that introduced extra trainable parameters, longer training times and more hyperparameters to tune. Perhaps, training should be done using better hardware or distributed computing environment to enhance efficiency of training. Finally, this research was mainly based on supervised learning; one might try using clustering or autoencoders or semi-supervised learning methods for deeper investigation. Another direction for further research is the extension of the base of areas, or the addition of other variables such as crime rates, education facilities or territories environmental impact before defining the final prediction setting.

References

1. Wang, Y., 2022. Population-land urbanization and comprehensive development evaluation of the Beijing-Tianjin-Hebei urban agglomeration. *Environmental Science and Pollution Research*, *29*(39), pp.59862-59871.
2. Barkham, R., Bokhari, S. and Saiz, A., 2022. Urban big data: city management and real estate markets. *Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities: Designing for Sustainability*, pp.177-209.
3. Deng, K.K., Chen, J., Lin, Z. and Yang, X., 2022. Differential selling strategies between investors and consumers: evidence from the Chinese housing market. *Journal of Real Estate Research*, *44*(1), pp.80-105.

4. Chen, D. and Li, R.Y.M., 2022. Predicting Housing Price in Beijing via Google and Microsoft AutoML. In *Current State of Art in Artificial Intelligence and Ubiquitous Cities* (pp. 105-115). Singapore: Springer Nature Singapore.
5. Xu, M. and Yang, Z., 2023, February. Research on the Influencing Factors Affecting Beijing House Prices Using Linear Regression Model. In *International Conference on Business and Policy Studies* (pp. 411-424). Singapore: Springer Nature Singapore.
6. Adetunji, A.B., Akande, O.N., Ajala, F.A., Oyewo, O., Akande, Y.F. and Oluwadara, G., 2022. House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, pp.806-813.
7. Ho, W.K., Tang, B.S. and Wong, S.W., 2021. Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), pp.48-70.
8. Rawool, A.G., Rogye, D.V., Rane, S.G. and Bharadi, V.A., 2021. House price prediction using machine learning. *Int. J. Res. Appl. Sci. Eng. Technol*, *9*, pp.686-692.
9. Shehadeh, A., Alshboul, O., Al Mamlook, R.E. and Hamedat, O., 2021. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, p.103827.
10. Soltani, A., Heydari, M., Aghaei, F. and Pettit, C.J., 2022. Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities*, *131*, p.103941.
11. Lee, S.H., Kim, J.H. and Huh, J.H., 2021. Land price forecasting research by macro and micro factors and real estate market utilization plan research by landscape factors: Big data analysis approach. *Symmetry*, *13*(4), p.616.
12. Bao, W., Tao, R., Afzal, A. and Dördüncü, H., 2022. Real estate prices, inflation, and health outcomes: Evidence from developed economies. *Frontiers in public health*, *10*, p.851388.
13. Duca, J.V., Muellbauer, J. and Murphy, A., 2021. What drives house price cycles? International experience and policy issues. *Journal of Economic Literature*, *59*(3), pp.773-864.
14. Barron, K., Kung, E. and Proserpio, D., 2021. The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, *40*(1), pp.23-47.
15. Duan, J., Tian, G., Yang, L. and Zhou, T., 2021. Addressing the macroeconomic and hedonic determinants of housing prices in Beijing Metropolitan Area, China. *Habitat International*, *113*, p.102374.
16. Xu, X. and Zhang, Y., 2022. Residential housing price index forecasting via neural networks. *Neural Computing and Applications*, *34*(17), pp.14763-14776.
17. Peng, H., Li, J., Wang, Z., Yang, R., Liu, M., Zhang, M., Philip, S.Y. and He, L., 2021. Lifelong property price prediction: A case study for the toronto real estate market. *IEEE Transactions on Knowledge and Data Engineering*, *35*(3), pp.2765-2780.