Predicting real estate prices use machine learning

1. Introduction

1.1 Problem

In recent years, the rapid development of artificial intelligence (AI) and machine learning (ML) technologies has created new opportunities for predicting and analysing the real estate market. With its complex and constantly fluctuating nature, this market requires more accurate forecasting tools to support investment and transaction decision-making. Big data plays a crucial role in providing rich and diverse information on the value, conditions, and trends of real estate. By combining big data with machine learning algorithms, we can delve deeper into the factors affecting property prices, thereby enabling more accurate predictions of future prices.

Real estate is a sector characterised by high uncertainty and randomness. Factors such as location, infrastructure, urban planning, and market demand are constantly shifting, creating volatility in asset values. Therefore, financial experts rely on forecasting and analytical models to maximise returns from real estate investments.

For urban designers, understanding forecasting models regarding future trends in infrastructure development, spatial planning, and construction design is extremely important. This knowledge helps them make strategic decisions, enhance leadership, and optimise decision-making processes. Grasping future trends and community needs supports them in developing effective and suitable solutions.

To accurately predict real estate costs and future development patterns, it is essential to gather a large amount of information on the market, planning, demographics, and economy. In-depth and reliable analyses will benefit investors, planners, and designers, helping them make more strategic and informed decisions.

1.2. Related Work

The prediction of real estate prices has been a topic of considerable research due to its significant impact on both individuals and the economy. A variety of machine learning algorithms have been employed to predict housing prices based on historical data and a range of features such as location, size, and amenities. This

section reviews previous research and compares the methodologies used in house price prediction.

Vyas and Sharma (2023) developed an algorithm leveraging linear regression to predict real estate prices with high accuracy. They experimented with several algorithms including Random Forest and Naive Bayes, concluding that linear regression was the most effective, yielding an accuracy of 85%. Their system used Python and the Scikit-learn library for implementation and Flask for the user interface. The model predicted property prices based on various features like the number of bedrooms, bathrooms, and square footage(An Algorithm to Predict…).

In a similar study, Manasa et al. (2020) explored regression techniques such as Lasso, Ridge regression, Support Vector Regression (SVR), and XGBoost to predict house prices in Bengaluru. The authors carried out a comparative study between these models using error metrics such as Root Mean Squared Error (RMSE) and R-Squared values. Their analysis revealed that XGBoost and SVR produced the most accurate results for the dataset, while linear regression served as a baseline. This study highlighted the importance of regularisation techniques in improving model performance by preventing overfitting(Machine Learning based …).

Other works in the field have focused on alternative methods like ensemble learning and feature extraction. For example, Pow et al. (2014) applied a combination of K-Nearest Neighbors (KNN) and Random Forest in an ensemble approach, which outperformed individual models with the least prediction error. Moreover, Wu and Yang (2017) explored the impact of feature extraction methods on the performance of Support Vector Regression (SVR), noting that proper feature selection can significantly reduce prediction error(Machine Learning based …).

Advanced models, such as neural networks, have also been applied to this domain. Limsombunchai (2004) compared a hedonic pricing model with an artificial neural network (ANN) for predicting house prices, finding that the ANN produced superior results with lower error rates. Tay and Ho (1992) also demonstrated that neural networks outperformed traditional regression models in predicting real estate prices (Machine Learning based …).

In conclusion, while linear regression remains a widely used and simple model for house price prediction, more complex models like SVR, XGBoost, and neural networks have shown potential in improving prediction accuracy. The use of regularisation techniques such as Lasso and Ridge further enhances model robustness by addressing multicollinearity and preventing overfitting. Future research should focus on combining these advanced techniques with real-time data to develop more accurate and scalable solutions for house price prediction.

1.3. Contribution

The project explores real estate-specific features that significantly impact home prices, such as property size, location, number of bedrooms, orientation, and legal status. Through rigorous feature engineering, it seeks to capture the nuances of real estate valuation, like the influence of proximity to urban centres, district-level characteristics, and neighbourhood trends.

Beyond price prediction, the analysis uncovers relationships between various property attributes and pricing trends. This provides a more comprehensive understanding of how factors like home orientation, frontage size, and accessibility impact property value over time.Investors can better assess the potential appreciation of properties in specific areas.Homebuyers gain insights into likely future prices, helping them make informed purchasing decisions.

2. Data

2.1. Data Overview

The dataset consists of **23,690 entries** and **15 columns**, providing detailed information about various real estate listings. Key attributes include Địa chỉ (address), which specifies the street and neighbourhood of the property, Phường/Xã/Thị trấn (ward or town), and Quận/Huyện (district). Property specifics like Diện tích (area in m^2), Mức giá (total price in billion VND), and Giá tiền (price per m^2 in million VND) are also provided. Additional property details include Mặt tiền (frontage width), Đường vào (access road width), Hướng nhà (house orientation), and Hướng ban công (balcony orientation). The dataset further captures structural information such as Số tầng (number of floors), Số phòng ngủ (number of bedrooms), and Số toilet (number of toilets). Legal and furnishing information is also available, with Pháp lý (legal status) indicating ownership documentation (e.g., "Đã có sổ đỏ") and Nội thất (interior furnishing status) describing the property's level of furnishing.

2.2. Data Collection

This dataset was collected from property listings on a Vietnamese real estate platform, specifically focusing on properties in Hanoi. Given its structure, the data

was likely gathered through web scraping, capturing a snapshot of listings available in October 2024. The source website, Bất động sản Việt Nam, provided extensive details for each property. To retrieve and compile this information, a web scraping tool utilizing BeautifulSoup was used, ensuring accurate and comprehensive data extraction.

2.3. Data Filtering and Split

To ensure data quality and relevance, several data filtering steps were applied. First, missing values in essential columns like "Mức giá" (total price), "Số phòng ngủ" (number of bedrooms), and "Giá tiền" (price per square meter) were handled by excluding listings with missing data, as these fields are critical for accurate analysis. Next, outliers in "Mức giá" and "Diện tích" (property area) were examined and treated to prevent extreme values from skewing results. Finally, columns such as "Diện tích" and "Giá tiền" were standardized by adjusting data types, including the removal of units, to ensure consistency across numerical values. After filtering, the data should be split into training and test sets, typically in an 80-20 or 70-30 ratio, to train predictive models and assess performance. This split helps ensure that each subset maintains similar distributions across key variables, like price and location.

3. Methodology

3.1. Data Reduction and Transformation

The initial dataset contains various features that describe property listings, some of which may be irrelevant or redundant. To enhance model efficiency and accuracy

3.2. Neural Network

-The prediction function of Random Forest can be represented as follows:
Neural Network learns complex relationships between house features (e.g. area, location, number of rooms, year built) and house value by adjusting the weights in the network.

-In the problem of predicting house prices, Neural Network can be used:

+ L2 regularization to reduce overfitting.
+ Dropout to randomly skip some neurons during training, helping the model learn better with noisy data

The prediction function of a Neural Network can be represented as:

$$\hat{y} = f(x) = f_L(W_L \cdot f_{L-1}(W_{L-1} \cdot ...f_1(W_1 \cdot x + b_1) + b_{L-1}) + b_L)$$

$\hat{y}$ is the predicted house price.

L is the number of layers in the network.

f is the activation function, such as ReLU in hidden layers and not applied in the output layer (continuous value regression).

W_i is the weight of layer i.

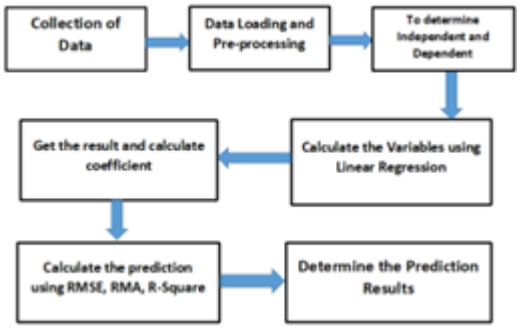
b_i is the bias of layer i.

Neural Networks are capable of modelling nonlinear relationships between influencing factors and house prices and allow for the incorporation of complex features such as location, area, number of rooms... to make more accurate predictions than simple linear models.

Comparison with Linear Regression: Linear Regression is simpler and easier to understand, but Neural Network is suitable for big data and strong nonlinear relationships. Neural Networks have better generalization capabilities for complex problems but require more computation.

3.3. Linear Regression

Linear Regression Models



Model Formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_x x_n + \epsilon$$

Trong đó:

- y is dependent variable
- $□_1, □_2, \ldots, □_□$ are the input features
- β_0 is the intercept (bias)
- $\beta_1, □_1, \ldots, □_□$ are the coefficients for each feature
- ϵ is the error term

The objective is to estimate the values of β in order to minimise prediction errors, typically using the Ordinary Least Squares (OLS) method. The OLS method minimises the sum of the squared differences between the actual values and the predicted values.

$$\text{Minimize } \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Trong đó:

- $□_□$ represents the observed value of the dependent variable for the i-th data point.
- $\hat{□}_□$ denotes the predicted value of the dependent variable for the i-th data point, calculated using the regression line.
- □ refers to the total number of observations.

To assess the performance of the linear regression model, the following metrics can be used:

1. R-squared (R^2):
    - Definition: This is the proportion of the variance in the dependent variable explained by the independent variables in the model. It ranges from 0 to 1.
    - Calculation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \in [0,1]$$

- A value of R^2 close to 1 indicates that the model explains the variability of house prices well.

2. RMSE (Root Mean Squared Error):
   - Definition: This is the square root of the average of the squared differences between predicted and actual values. RMSE indicates the average deviation between predicted values and actual values.
   - Calculation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \in [0, +\infty)$$

   - A lower RMSE value signifies greater accuracy of the model.

3. MAE (Mean Absolute Error):
   - Definition: MAE is the average of the absolute differences between predicted and actual values.
   - Calculation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \in [0, +\infty)$$

   - A lower MAE value indicates higher accuracy of the model.

3.3 Random Forest

Random Forest is a powerful machine learning model that uses multiple decision trees, by using a forest of many independent decision trees and averaging their results Random Forest helps to reduce overfitting and increase accuracy.

The prediction function of Random Forest can be represented as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

$\hat{y}$ is the predicted house price.

N is the number of trees in the forest.

$T_i(x)$ is the prediction from the i-th decision tree for a sample x.

Comparison with Linear Regression: Linear Regression works well with linear relationships, while Random Forest can capture more complex nonlinear relationships.

3.4. Conclusion

Linear Regression is a reasonable choice for a house price prediction problem when you want a model that is simple, easy to understand, and capable of explaining the linear relationship between characteristics and house value. Linear Regression is appropriate if your data has a roughly linear relationship, is not too complex, and has little noise.