Northeastern University

INFO 6150 Data Science Engineering Methods and Tools

Final Project Report - Fall 2024

Statistical Analysis and Prediction of Real Estate Prices

by
Prachi Navale

supervised by
Prof. Hong Pan

Abstract

What makes a house truly valuable? Is it its age, its proximity to public transport, or the bustling convenience stores just around the corner? This project dives into these questions, unraveling the mysteries of real estate valuation using data-driven insights. By analyzing a rich dataset from the UC Irvine Machine Learning Repository [1], we uncover the key factors that influence housing prices, from the age of a property to its distance from the nearest MRT station. Using machine learning models like Linear Regression and Random Forest, we explored the data to reveal not just numbers but stories — of neighborhoods, infrastructure, and market trends. The results? A Random Forest model that predicts housing prices with remarkable accuracy, shedding light on the complex dynamics of real estate markets. Beyond the numbers, this study offers a fresh perspective for urban planners, investors, and homeowners alike: the value of a home lies not just in its walls, but in the world it's connected to.

Contents

1 Introduction

Real estate valuation is more than just crunching numbers; it's about understanding the story behind every property. What makes one house worth more than another? Is it the bustling convenience stores nearby, its proximity to public transport, or the charm of its age? As cities grow and markets shift, traditional methods of determining property value often miss the bigger picture. That's where data and technology come in.

This project takes a deep dive into real estate valuation using machine learning, fueled by a dataset from the UC Irvine Machine Learning Repository [1]. The dataset captures a range of features like the age of a house, how close it is to the nearest MRT station, and even the number of convenience stores in the area. These attributes reveal the hidden factors that drive property prices, helping us make sense of the complex real estate market.

The goal of this project is simple but powerful: predict housing prices accurately and uncover the factors that matter most. Using Linear Regression [2] and Random Forest models [3], we analyzed the data to understand not just what the numbers say but what they mean. Random Forest emerged as the star, delivering highly accurate predictions and capturing the intricate relationships between location, amenities, and pricing.

This study isn't just about algorithms; it's about making better decisions. Whether you're an investor, a planner, or someone dreaming of their next home, these insights can guide smarter choices. By blending data and storytelling, this project bridges the gap between the art and science of real estate valuation.

2 Methods

2.1 Data Preparation

The first step in this analysis involved preparing the dataset for machine learning. The Real Estate Valuation dataset, sourced from the UC Irvine Machine Learning Repository, was loaded into Python for preprocessing. Key steps included:

- Ensured no missing values, excluded irrelevant features, and scaled numeric attributes using MinMaxScaler for consistent ranges
- Split the data into training (80%) and testing (20%) subsets for model training and evaluation

2.2 Exploratory Data Analysis (EDA)

EDA provided valuable insights into the factors influencing housing prices using tools like Pandas [4], Matplotlib [5], and Seaborn [6]:

- Visualized data distributions using histograms and box plots
- Examined feature relationships using scatter plots, heatmaps, and correlation analysis
- Conducted geospatial analysis to explore price clustering by location

2.3 Machine Learning Models

To predict housing prices and analyze feature importance, two machine learning models were implemented:

- Implemented **Linear Regression** as a baseline model to capture linear relationships between features and housing prices
- Used **Random Forest Regressor** to model complex, non-linear patterns in the data

2.4 Evaluation Metrics

The performance of the models was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE):** To measure the average error magnitude, penalizing larger errors more heavily
- **Mean Absolute Error (MAE):** To evaluate the average error magnitude, providing a more intuitive measure of model accuracy
- **R-squared ($\mathbf{R^2}$):** To determine the proportion of variance in house prices explained by the model, offering a measure of goodness-of-fit

3 Results

3.1 EDA Outcomes

The exploratory data analysis (EDA) revealed several key patterns and insights into the dataset:

- **Distribution of House Prices:** A histogram showed that house prices per unit area were slightly skewed to the right, indicating a concentration of properties in the mid-to-low price range with a few outliers at the higher end.

- **Correlation with Distance to MRT Stations:** A scatter plot and correlation matrix highlighted a strong negative relationship between house prices and distance to MRT stations ($r = -0.67$), suggesting that proximity to public transport significantly increases property value.

- **Effect of Convenience Stores:** Houses located near more convenience stores tended to have higher unit area prices. Box plots comparing properties with varying numbers of nearby stores illustrated this trend.

- **Geographical Insights:** A scatter plot of latitude and longitude revealed clusters of high-priced houses in specific regions, indicating the influence of desirable neighborhoods on valuation.

- **House Age Analysis:** Surprisingly, house age had a mixed impact. While older houses generally trended toward lower prices, some maintained high values due to their location or unique attributes.

3.2 Model Performance

To evaluate the ability of machine learning models to predict housing prices, the following metrics were calculated on the train as well as test dataset for each model:

Model	**RMSE**	**MAE**	**R-squared (R^2)**
Linear Regression	0.079	0.053	0.59
Random Forest	0.027	0.017	0.95

Table 1: Model Performance Metrics for Train Dataset

Model	**RMSE**	**MAE**	**R-squared (R^2)**
Linear Regression	0.065	0.048	0.69
Random Forest	0.052	0.035	0.80

Table 2: Model Performance Metrics for Test Dataset

- **Linear Regression:** The baseline model provided reasonable predictions but struggled with non-linear relationships, as evident from its lower R^2 score and higher error rates

- **Random Forest:** This model significantly outperformed Linear Regression, with a much lower RMSE and MAE, and a high R^2 value, indicating its strength in capturing complex patterns in the data

4 Discussion

The results of this project provide valuable insights into real estate valuation, highlighting both the strengths of machine learning models and the challenges of working with complex datasets.

Model Performance

The Random Forest model demonstrated strong predictive capabilities, outperforming Linear Regression across all evaluation metrics. With an R^2 of 0.80, it effectively captured the non-linear relationships between house prices and features like proximity to MRT stations and neighborhood amenities. This model's ability to handle interactions and non-linear dependencies makes it a superior choice for real estate valuation, where such complexities are common.

Linear Regression, while easier to interpret, struggled with the dataset's intricacies. Its performance, though reasonable (R^2 of 0.69), was hindered by its assumption of linearity, making it less suitable for capturing nuanced relationships. This highlights the importance of selecting the right model for the problem at hand, balancing accuracy with interpret-ability.

4.1 Implications of Results

The findings reinforce well-known principles in real estate:

- **Proximity to Public Transport:** Properties closer to MRT stations command higher prices, emphasizing the importance of accessibility in urban planning.
- **Neighborhood Amenities:** The number of nearby convenience stores is a significant driver of property value, underlining the role of infrastructure in shaping housing demand.
- **Location Matters:** Geospatial clustering of high-priced properties reflects the enduring influence of desirable neighborhoods on real estate markets.

These insights have practical applications for various stakeholders:

- **Urban Planners:** Use such models to assess how changes in infrastructure impact property values.

- **Investors:** Identify high-value areas by analyzing patterns in location-based features.
- **Policy Makers:** Prioritize public transport and amenity development to boost property values and livability.

4.2 Limitations and Challenges

Despite the success of the models, several challenges were encountered:

- **Feature Selection:** While the dataset provided key attributes, additional factors like crime rates, school quality, and economic conditions could enhance model performance.
- **Data Quality:** The dataset lacked categorical features (e.g., property type) and temporal trends, limiting its scope for broader generalizations.
- **Overfitting Risk:** The Random Forest model, while highly accurate, is prone to overfitting on smaller datasets, necessitating careful parameter tuning.
- **Scalability:** Applying this approach to larger, more diverse datasets may require additional preprocessing and computational resources.

4.3 Future Directions

To address these limitations, future work could explore:

- **Incorporating Additional Datasets:** Use datasets with richer feature sets, such as demographic and economic data.
- **Experimenting with Advanced Models:** Try Gradient Boosting Machines or Neural Networks for improved predictions.
- **Conducting Temporal Analyses:** Account for market trends and seasonal variations in the data.

5 Conclusion

This project successfully identified the key factors influencing real estate prices and evaluated machine learning models to predict housing values. Proximity to MRT stations and neighborhood amenities, such as convenience stores, emerged as the most significant drivers of property prices,

while high-value properties were clustered in specific desirable neighborhoods. Among the models tested, Random Forest outperformed Linear Regression, achieving an R^2 of 0.80 compared to 0.69, effectively capturing the non-linear relationships in the data. However, the analysis could benefit from incorporating additional features like crime rates and school quality, as well as temporal data to account for market trends over time. Future work could explore advanced models like Gradient Boosting or Neural Networks and scale the approach to larger datasets. This study lays the groundwork for data-driven real estate valuation, offering practical insights for urban planners, investors, and policymakers.

References

[1] UCI Machine Learning Repository, "Real estate valuation data set," n.d., retrieved from https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set.

[2] S. Weisberg, *Applied linear regression*, 3rd ed. Wiley-Interscience, 2005.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 56–61, pandas Library.

[5] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, matplotlib Library.

[6] M. Waskom *et al.*, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021, seaborn Library.

Youtube Project Run Demo Link: Project Run Demo