Analysis and Prediction of Apartment Prices in Ulaanbaatar, Mongolia

Jargalsaikhan Namuun[1], Yangin Yoon*
Department of Data Science, Seoul National University of Science & Technology
[1]namuun@ds.seoultech.ac.kr, *ben.yangin.yoon@seoultech.ac.kr

본 연구는 2025년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥회의 지원을 받아 수행되었음 (P0017123 , 2025년 산업혁신인재성장지원사업)

Research Overview

Background:

- Ulaanbaatar’s real estate market is growing but lacks transparency and available data.
- Apartment price prediction is important for investors, buyers, and policymakers.

Research Objective:

- To predict apartment prices accurately using machine learning models.
- To explore key factors influencing apartment prices.

Problem Statement:

- Lack of open-access datasets.
- Limited machine learning applications in Mongolia’s real estate market.

Research Trends & Gap

Domestic and International Research:

Year	Authors	Country/Region	Used methodology	Contribution
2017	Alfiyatin et al.	Indonesia	Regression + Particle Swarm Optimization	- Tuned PSO improved RMSE - Effective for local Indonesian housing data
2019	Amarbayan Altangerel	Mongolia	XGBoost, RF, LR	- Prior attempt using ML (non-academic, Medium post) - The best performing model is XGBoost
2020	Ahtesham et al.	Pakistan	Linear Regression, Random Forest, XGBoost	Applied ML to Karachi housing market, identified significant predictors.
2021	Erdenebat & Buyannemekh	Mongolia	Multiple Linear Regression	Used 51,396 apartment records to quantify feature impact.
2023	Dhar & Manikandan	India	SVM, Decision Tree, Random Forest, Linear Regression, ANN	Random Forest and ANN performed best among traditional models.
2024	Adzanoukpe	Ghana	CatBoost, XGBoost, RF, SVR, LR	CatBoost outperformed all.
2025	My research	Mongolia	Linear Regression, SVR, RF, XGBoost, CatBoost (with hyperparameter tuning)	Compared ML models, Random Forest performed best.

Gap Identification:

- No prior study has combined scraped apartment data with advanced ML model tuning.
- Little focus on feature importance analysis.

Model Pipeline

Define Research purpose & Scope	→ Set research objective	Previous studies	
Data Collection	→ Scrape data from Unegui.mn	Select key features	
Data Preprocessing	→ Handle missing values	One-hot encoding	Feature scaling
Model Selection & Training	→ LR, SVR, RF, XGBoost, CatBoost		
Hyperparameter Tuning	→ GridSearch	RandomizedSearch	Bayesian Optimization
Model Evaluation	→ Metrics: R-squared, MAE, RMSE	Train/Test split: 80/20	Compare results before/after tuning
Result	→ Feature importance	Best model	

Implementation and Preprocessing

Data Source:

- Platform: Unegui.mn - Mongolia's most widely used advertisement site.
- Method: Web scraping using Python.
- Final Dataset: 19,400+ apartment listings, was collected on 2025/03/19

Data Cleaning:

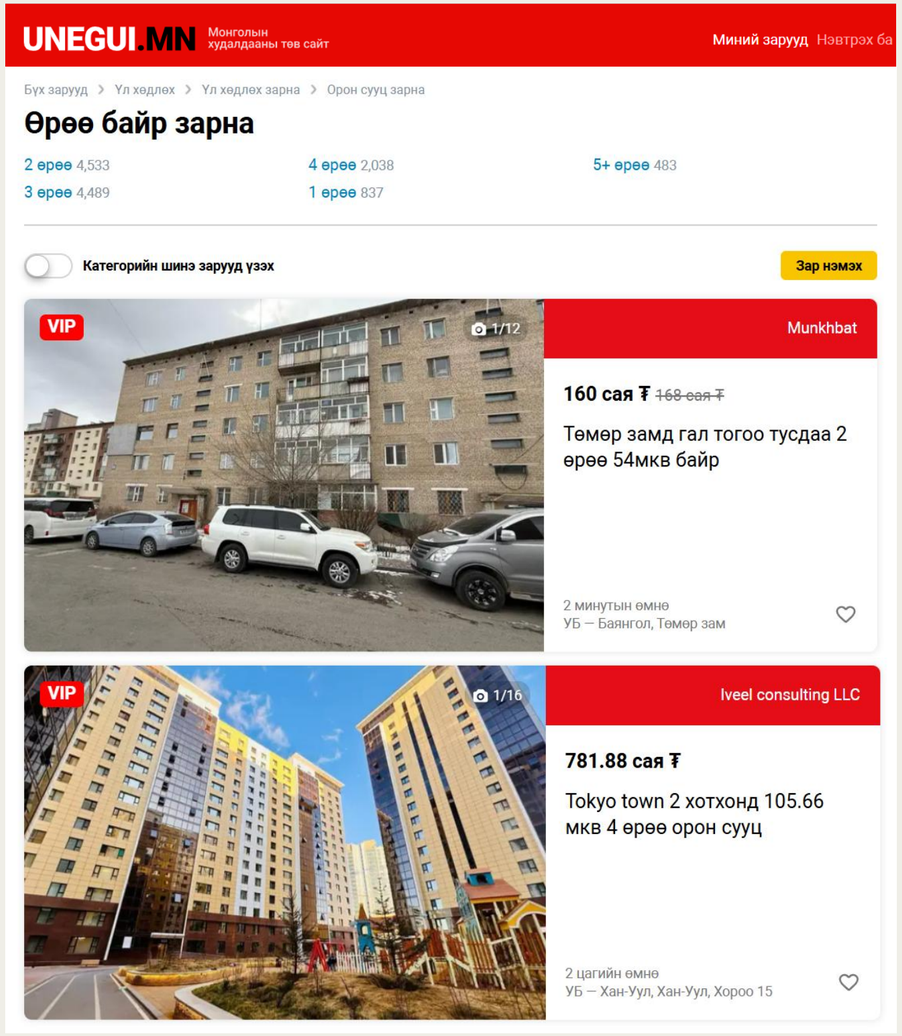
- Removed duplicate or incomplete entries.
- Converted object columns into numerical types.

Model Training:

- Train-test split ratio: 80:20.
- Used models: Linear Regression, SVR, Random Forest, XGBoost, CatBoost.

Hyperparameter Tuning:

- Tried GridSearchCV, RandomizedSearchCV, and Bayesian Optimization.
- Compared performance using R^2, MSE, RMSE, MAE.
- Final model selection was based on the best metric scores

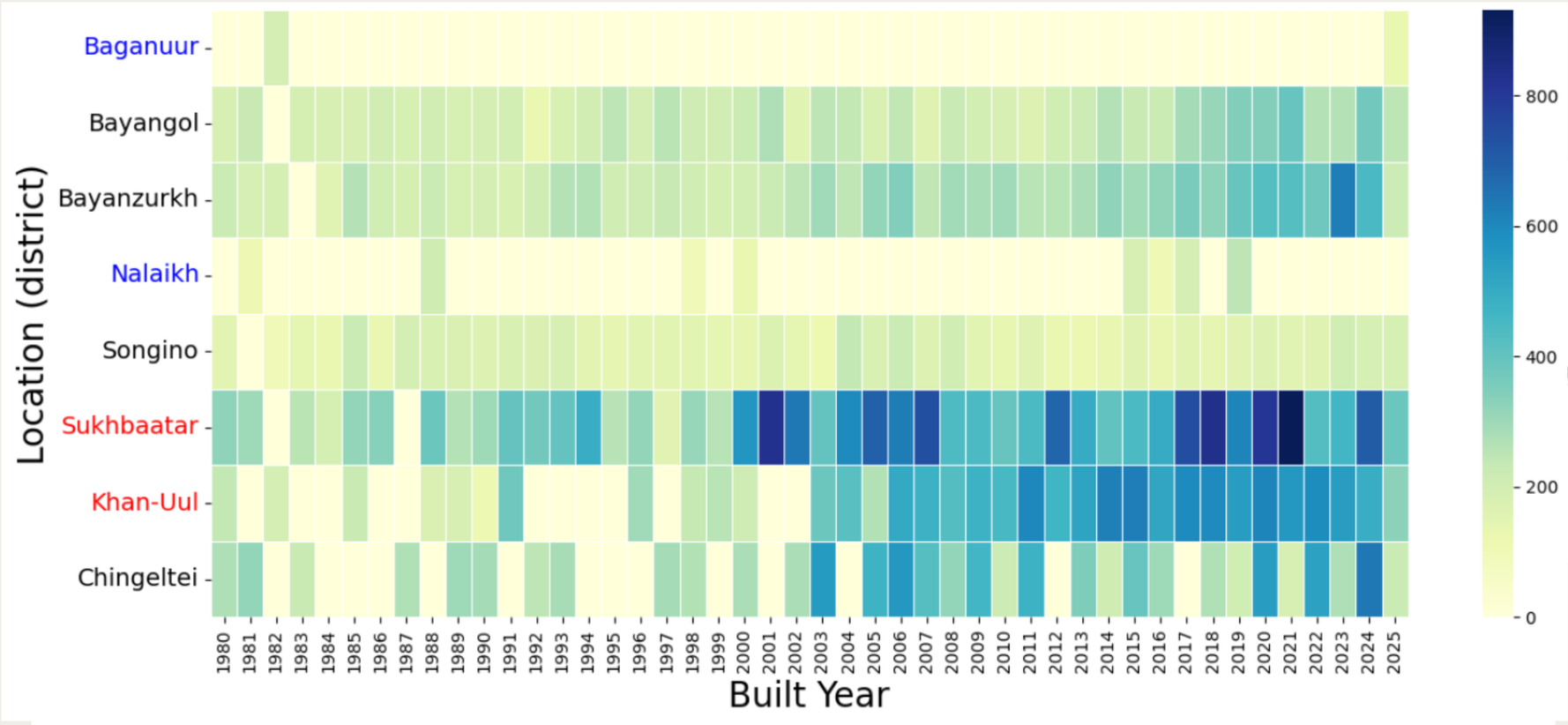


Model Input Variables

Category	Feature names
Structural features	total_floor, located_floor, size_m2, number_of_rooms, number_of_windows
Building condition	built_year, construction_progress, elevator
Amenities	floor_type, balcony, garage, window, door
Location info	location
Transaction info	payment_term, posted_date
Target variable	price

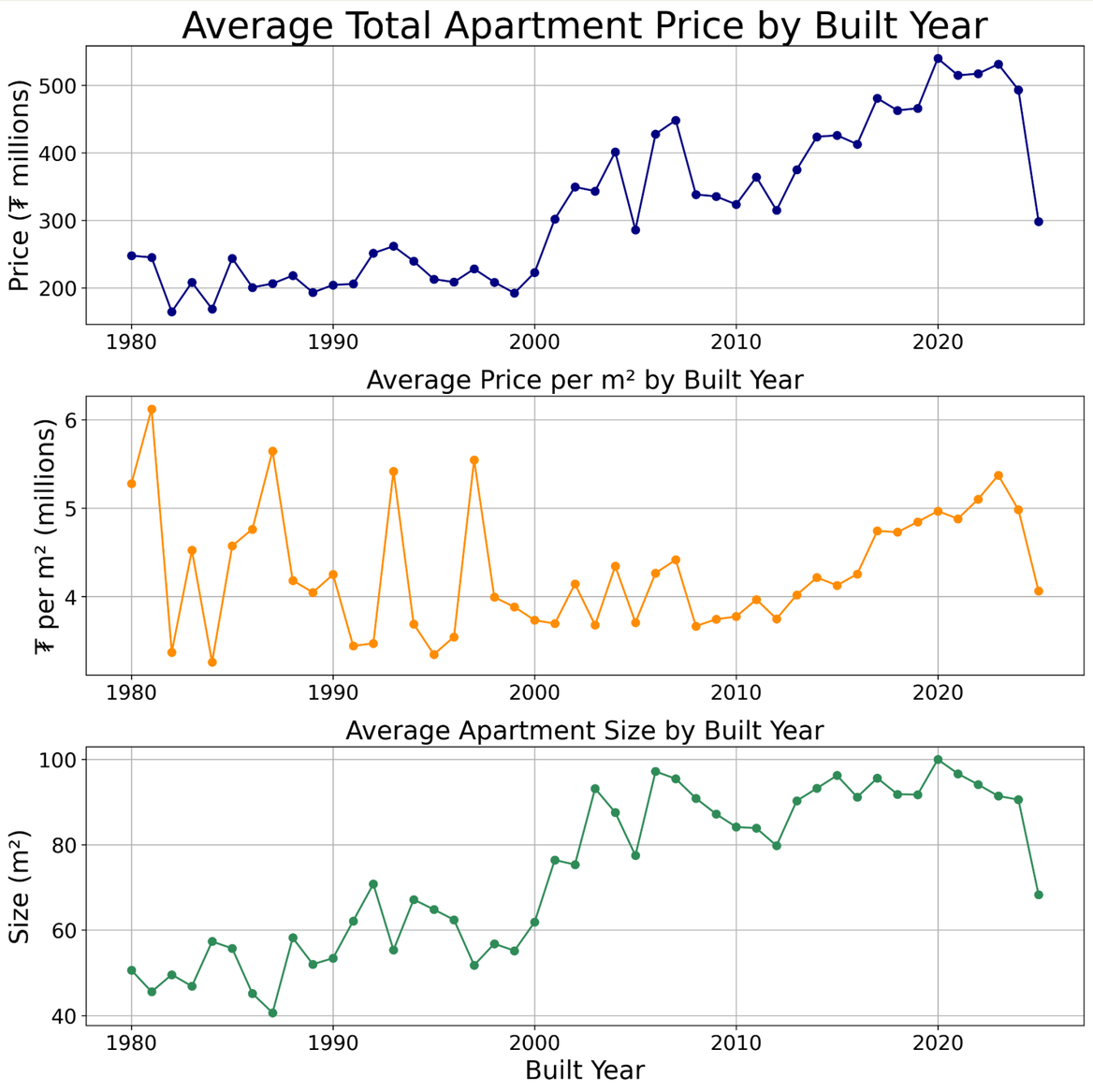
- A total of 16 features were used to train the model, including structural characteristics, amenities, building condition, and location.
- All features were fully non-null, cleaned and encoded as numerical/categorical.
- The target variable is apartment price (in million ₹)

Apartment Price Trends by District and Year



- Newer properties in city center districts (Sukhbaatar and Khan-Uul) are significantly more expensive, while those in outer districts (Nalaikh and Baganuur) remain relatively affordable.
- The slight drop or inconsistency in prices for recent years (2024–2025) may reflect limited data availability.

Key EDA Findings



- While newer apartments appear more expensive in total price, the **price per m² remains relatively stable** across decades.
- Despite being built 20–40 years ago, older apartments show comparable or even higher price per m² than many newer ones showing **older apartments retain strong value.**
- There's a clear trend toward increased apartment size over time. Apartments built after 2010 are **30–50% larger** than those from the 1980s–1990s.

Model Performance Before Tuning

Model	R^2	MSE	RMSE	MAE
Linear Regression	0.599	0.119	0.345	0.275
Random Forest Regressor	0.964	0.011	0.103	0.053
Support Vector Regressor	0.832	0.050	0.224	0.163
XGBoost	0.920	0.024	0.154	0.109
CatBoost	0.898	0.030	0.174	0.128

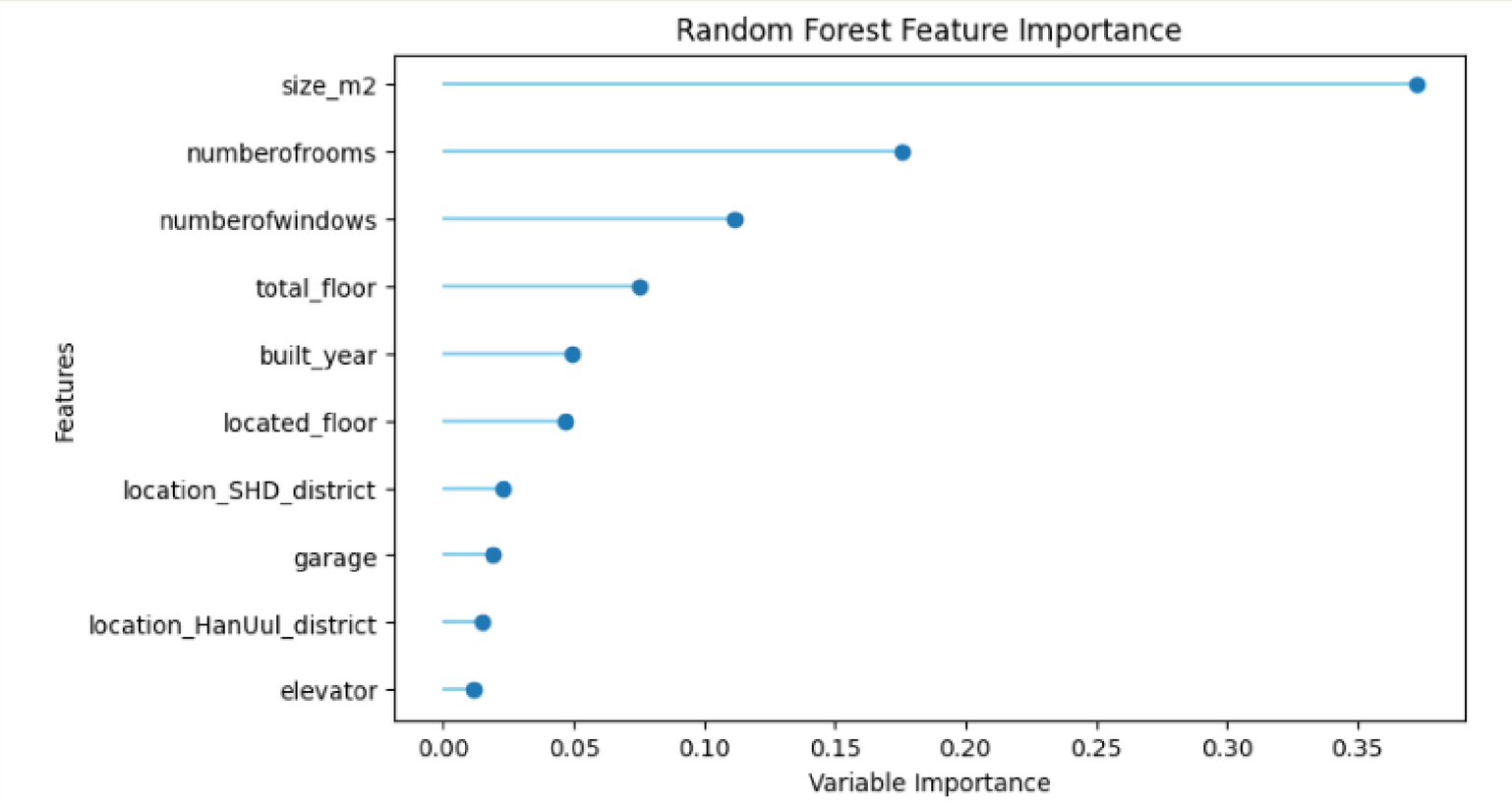
- All five models were first trained using default parameters.
- Random Forest Regressor showed the best performance with the highest R^2 (0.964) and lowest error metrics.
- Linear Regression performed the worst.

After Hyperparameter Tuning (Best Scores)

Model	Method	R^2	MSE	RMSE	MAE
Linear Regression	-	0.599	0.119	0.345	0.275
Random Forest Regressor	BO	**0.973**	**0.008**	**0.090**	**0.028**
Support Vector Regressor	GridSearch	0.914	0.025	0.159	0.114
XGBoost	GridSearch	0.962	0.011	0.106	0.060
CatBoost	BO	0.967	0.010	0.099	0.051

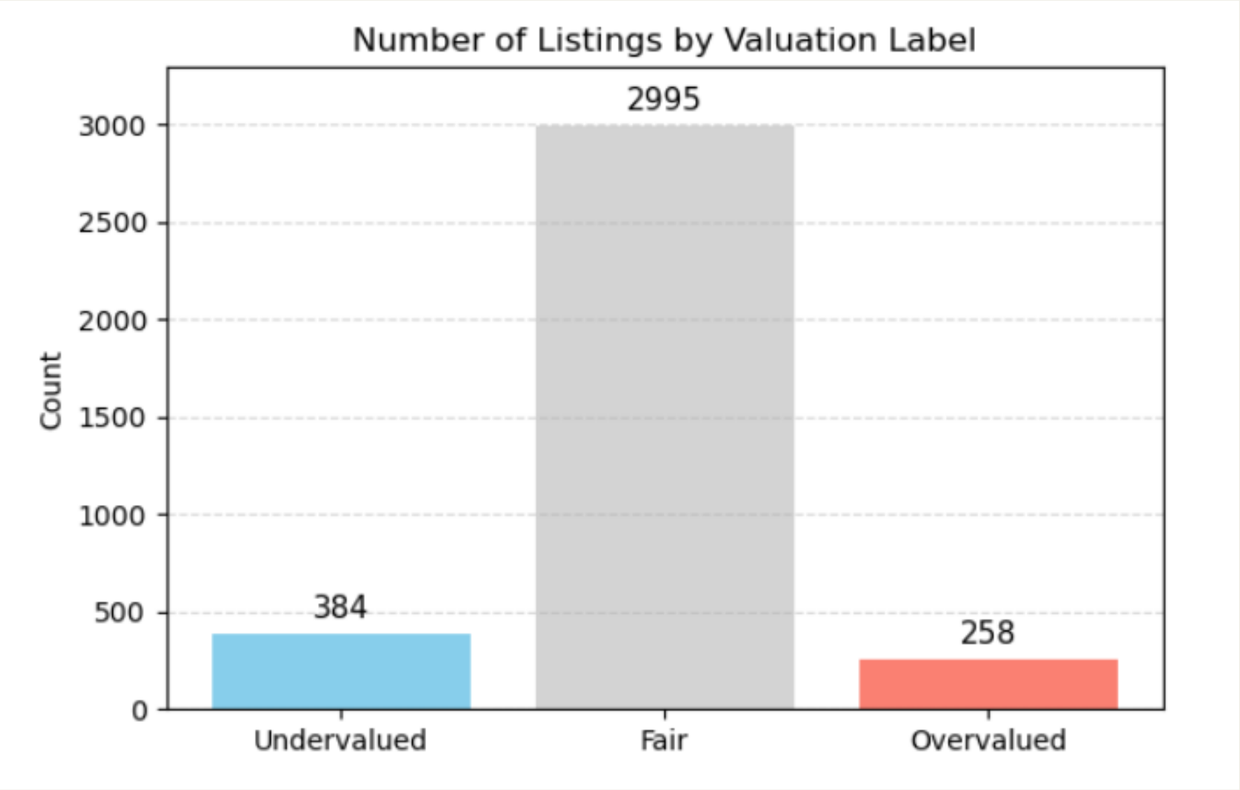
- To improve model performance, three hyperparameter tuning methods were applied to all models: GridSearch, RandomizedSearch and Bayesian Optimization (BO).
- For each model, the best-performing tuning result was recorded.
- Among all, **Random Forest Regressor tuned by Bayesian Optimization** achieved the highest R^2 (0.973) and lowest MAE (0.028). This shows that tuning significantly improves prediction accuracy, and Bayesian Optimization is especially effective for this dataset.

Feature Importance Analysis



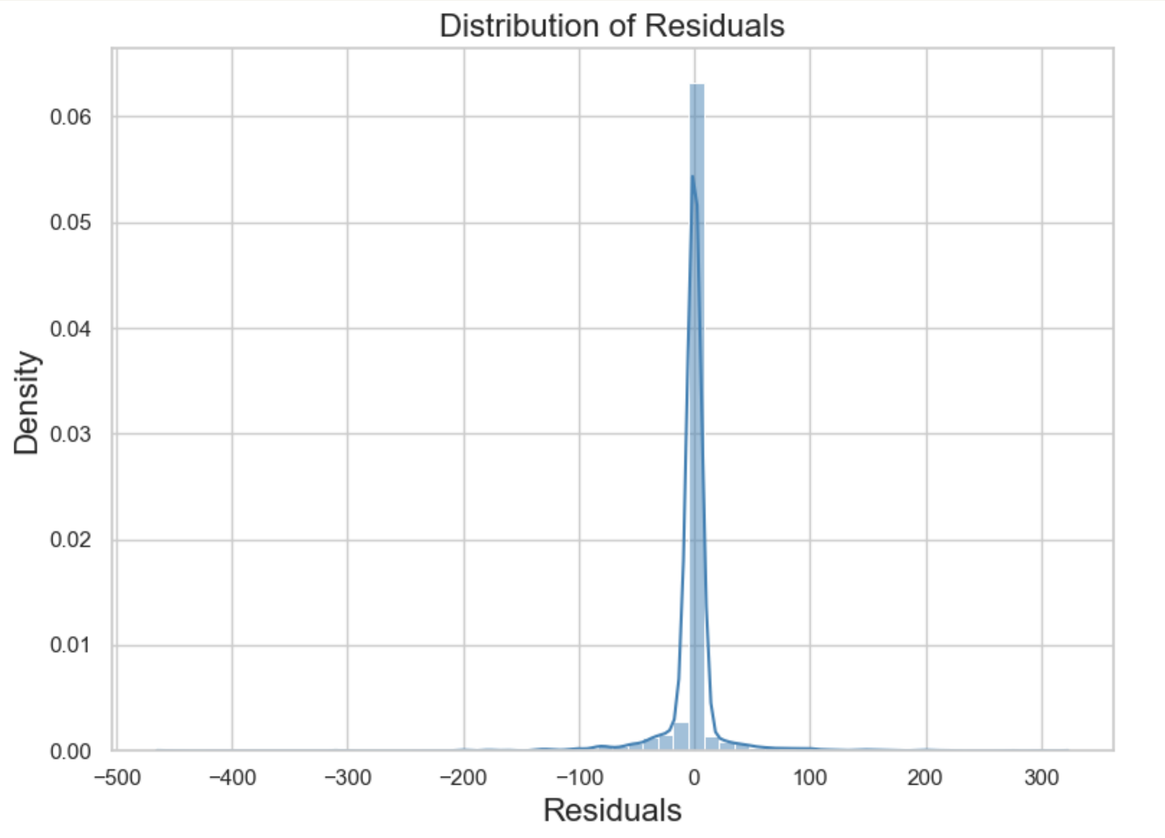
- **Size (m²)** is the most influential factor in determining apartment prices.
- Interior features (rooms, windows) come next.
- Location variables and infrastructure features like elevator and garage have minimal impact in the current model.

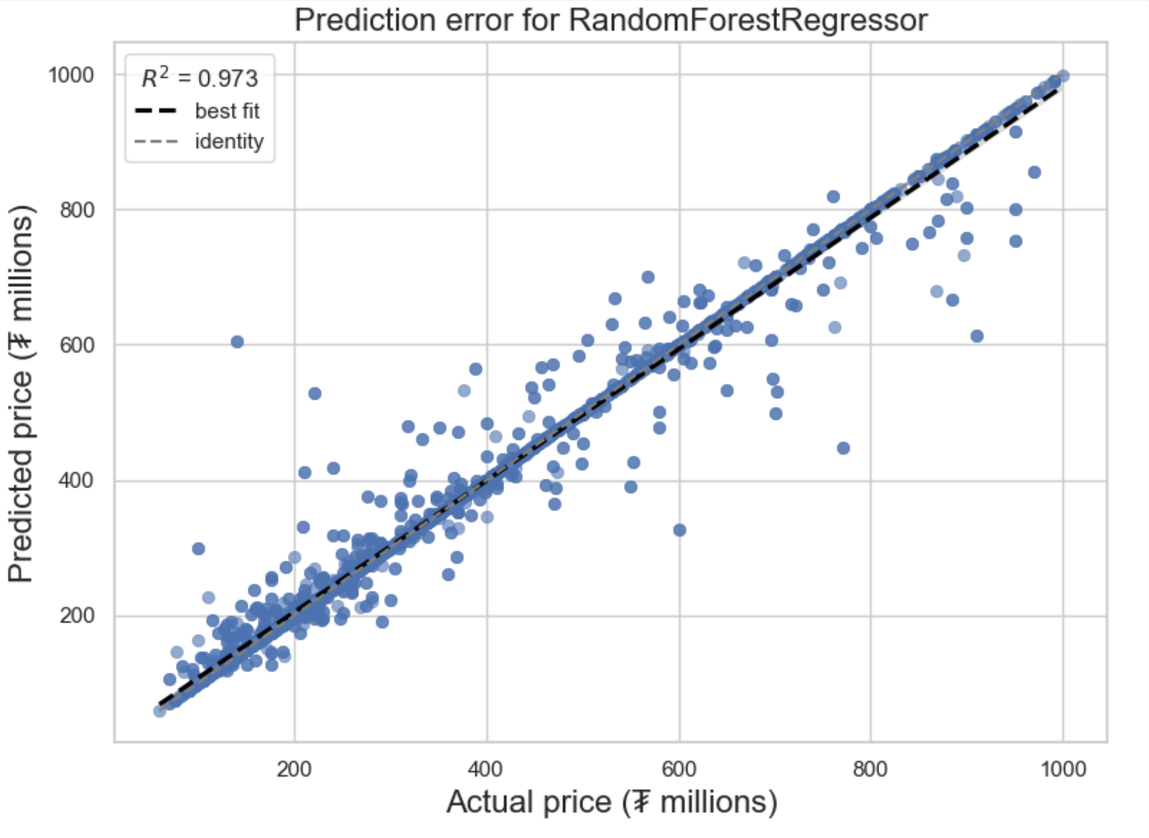
Model-Based Valuation Labels



- Using a ₹5 million price gap threshold, most apartment listings were found to be **fairly priced**.
- However, 384 listings (10%) were **undervalued**, meaning their actual listed price is significantly below the model's predicted value. This may signal **potential investment opportunities**.
- On the other hand, 258 listings (7%) appear **overpriced**, suggesting market inefficiencies or aggressive seller pricing.

Visualization of Results





- Most residuals are centered tightly around 0, indicating low bias and consistent model performance.
- The model achieved an R^2 of 0.973, with predicted prices closely aligned to actual prices along the identity line.

Limitations and Future work

- Although over 19,000 listings were collected, all data were sourced from a single platform (Unegui.mn), which may not fully represent the entire real estate market. Future studies can integrate data from **multiple sources**, including real estate agencies, or actual transaction datasets.
- The models were trained on **asking prices**, not actual transaction prices. This limits the model's ability to reflect true market behavior. Comparing predicted prices with **actual sales data** can help improve the model's real-world accuracy.
- While **Random Forest** showed the highest accuracy, it lacks interpretability compared to simpler models like linear regression. To improve transparency, future research may apply **model-agnostic explainability methods** such as SHAP or LIME.

Conclusion and Implications

- This study demonstrated that machine learning models, especially Random Forest with Bayesian Optimization can accurately predict apartment prices in Ulaanbaatar, achieving an R^2 of 0.973.
- Size (m^2) and interior features were found to be the most influential factors.
- Older apartments which were built 20–40 years ago retain strong value per m^2, suggesting long-term investment potential even in aging buildings.
- A model-based valuation approach revealed that around 10% of listings are significantly undervalued, highlighting potential investment opportunities. On the contrary, overpriced listings suggest areas for buyer caution.
- These insights can support data-driven decision-making for real estate investors, developers, and policymakers aiming to improve housing affordability and market transparency in Mongolia.

Reference

1. Adyan Nur Alfiyatin, Adyan Nur Alfiyatin. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications 8(10), 323–326.* https://doi.org/10.14569/IJACSA.2017.081042
2. Amarbayan Altangerel. (2019). Predicting Ulaanbaatar's apartment price https://medium.com/@weatheranchor_43165/predicting-ulaanbaatars-apartment-price-c3dcccbaee57#:~:text=Where%20should%20we%20look%20for,our%20data
3. Maida Ahtesham, Narmeen Zakaria Bawany, Kiran Fatima. (2020). House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan. *2020 21st International Arab Conference on Information Technology (ACIT).* https://doi.org/10.1109/ACIT50332.2020.9300074
4. Erdenebat M, Buyannemekh B. (2021). The Effect of Varying Characteristics of Residential Apartments on Their Value. *Journal of Business and Innovation, 7(2), 67–80.* https://journal.num.edu.mn/BusinessAndInnovation/article/view/1747
5. Tanmoy Dhar, Manikandan P. (2023). A Literature Review on Using Machine Learning Algorithm to Predict House Prices. *International Research Journal on Advanced Science Hub*, 5(05), 132-137. http://dx.doi.org/10.47392/irjash.2023.S017
6. Adzanoukpe, P. (2025). Predicting House Rental Prices in Ghana Using Machine Learning. *arXiv preprint arXiv:2501.06241.* https://doi.org/10.48550/arXiv.2501.06241