Real Estate Price Prediction and Feature Analysis:

A Comparative Study of Housing Models in U.S. and Malaysian Markets

Vasilov Dmitrii, Department of Computer Science, *Universiti Putra Malaysia,* Serdang, Selangor, Malaysia.
Email: DMVasilov@yandex.ru

This study presents a comprehensive machine learning approach to residential real estate price prediction across two distinct markets: Ames, Iowa (USA) and Malaysia. We implement and compare three regression models (Lasso, Random Forest, and XGBoost) and evaluate their performance through extensive experimentation. Results demonstrate that XGBoost achieves superior performance with R² scores of 0.9067 for Ames and 0.4271 for Malaysia, highlighting the significant impact of data quality and feature engineering on model effectiveness. An interactive web application was developed to demonstrate practical deployment of the best-performing models.

Keywords: Real Estate, Machine Learning, XGBoost, Price Prediction, Regression Analysis, Web Application

I. Introduction

Real estate price prediction remains a critical challenge in financial markets, property valuation, and investment decision-making[1][2]. Traditional appraisal methods often rely on expert judgment and limited comparable sales, potentially missing complex non-linear relationships between property features and market values[3]. Machine learning techniques offer data-driven approaches to capture these intricate patterns.

This research investigates the application of three distinct machine learning algorithms - Lasso Regression, Random Forest, and XGBoost—to predict residential property prices in two geographically and economically diverse markets: Ames, Iowa (USA) and Malaysia. The Ames Housing Dataset (2006-2010, 2,930 properties) provides a mature, well-documented market with extensive feature engineering, while the Malaysia dataset (2025, 1,877 properties) represents an emerging market with distinct characteristics.

The primary contributions of this work include: (1) comprehensive comparative analysis of three regression algorithms across diverse markets, (2) feature importance analysis revealing market-specific pricing drivers, (3) empirical demonstration of data quality impact on model performance, and (4) development of an interactive web application for practical model deployment.

II. Related works

Numerous studies have explored machine learning applications in real estate valuation. Regression trees and ensemble methods have shown promising results for property price estimation[4]. Recent work has demonstrated XGBoost's effectiveness in capturing non-linear relationships in housing data[5][6].

Cross-market comparative studies remain limited, with most research focusing on single geographic regions. Our work addresses this gap by systematically comparing model performance across markets with different data characteristics, providing insights into generalization capabilities and data quality requirements.

III. Methodology

Datasets

Ames Housing Dataset: comprehensive collection of 2,930 residential properties sold in Ames, Iowa (2006-2010). Contains 79 explanatory variables describing property characteristics including physical attributes, location, quality ratings, and sale conditions. After preprocessing and feature engineering, 268 features were utilized for model training.

Malaysia Dataset: collection of 2,000 residential properties from various Malaysian states (2025). Contains location data, property size, rooms, bathrooms, parking, and furnishing status. After preprocessing, 2,179 features were created through extensive categorical encoding. Note: median price per square foot was excluded from training to prevent data leakage.

Data preprocessing

Data preprocessing consisted of multiple stages:

1. Missing value handling: domain-specific imputation strategies were applied. For Ames data, missing values in basement features indicated absence of basements (filled with 0). For Malaysia data, sparse missing values were removed.

2. Outlier detection: properties with extreme characteristics were identified and removed using domain knowledge and statistical methods.

3. Feature engineering:
- Ames: Created combined features (Total Area = GrLivArea + TotalBsmtSF), temporal features from year variables;
- Malaysia: One-hot encoding for categorical variables (Location, Property Type, Furnishing).

4. Feature scaling: StandardScaler applied to normalize continuous features and ensure equal contribution to distance-based algorithms.

Models

Lasso Regression: L1-regularized linear regression that performs feature selection by driving coefficients of irrelevant features to zero. Provides interpretable linear relationships.
- Ames: α = 1000, max_iter = 5000
- Malaysia: α = 5000, max_iter = 5000

Random Forest: ensemble method combining multiple decision trees with bootstrap aggregation and random feature selection. Captures non-linear relationships while reducing overfitting through averaging.
- Ames: 100 estimators, max_depth = 20
- Malaysia: 100 estimators, max_depth = 15

XGBoost: gradient boosting algorithm that builds trees sequentially, with each tree correcting errors of previous trees. Incorporates regularization and advanced optimization techniques.
- Ames: 200 estimators, max_depth = 7, learning_rate = 0.1
- Malaysia: 150 estimators, max_depth = 6, learning_rate = 0.15

Evaluation metrics

Models were evaluated using:
- R^2 Score: proportion of variance explained (0 to 1, higher is better);
- RMSE (Root Mean Squared Error): average prediction error magnitude in price units;
- MAE (Mean Absolute Error): average absolute prediction error.

Train/validation split: 80/20 with random_state=42 for reproducibility.

IV. Results

Ames Market performance

Model	**R^2 Score**	**RMSE**	**MAE**
XGBoost	0.9067	$18,632	$12,864
Random Forest	0.8891	$20,305	$14,035
Lasso	0.8284	$25,261	$14,652

XGBoost achieved the highest performance, explaining 90.67% of price variance with an average error of $18,632. Random Forest demonstrated competitive performance (88.91% R^2), while Lasso's linear assumptions limited performance (82.84% R^2).

Malaysia Market performance

Model	**R^2 Score**	**RMSE**	**MAE**
XGBoost	0.4271	RM 156,786	RM 114,488
Random Forest	0.3714	RM 164,222	RM 118,821
Lasso	0.3200	RM 170,805	RM 126,425

XGBoost again outperformed alternatives but with significantly lower R^2 (42.71%). The moderate performance across all models suggests inherent data limitations and greater market complexity.

Feature importance analysis

Ames Market - Top 5 features:
1. Overall Qual: 32.77%
2. Total Area: 9.76%
3. Garage Cars: 8.54%
4. Kitchen Qual_TA: 3.54%
5. Exter Qual_TA: 2.84%

Overall Quality dominates predictions, with physical size and garage capacity as secondary factors. Feature importance is distributed across multiple attributes.

Malaysia Market - Top 5 features:
1. Type Flat: 8.41%
2. State Selangor: 3.16%
3. State Kedah: 2.05%
4. Area Tebrau: 1.55%
5. Tenure Leasehold: 1.47%

The type of the housing and location features dominate Malaysia predictions, with areas showing highest importance. This indicates location-driven pricing in emerging markets.

Cross-market comparison

Factor	**Ames (USA)**	**Malaysia**
Best R^2	90.67%	42.71%
Prediction quality	Excellent	Moderate
Primary driver	Overall quality (32.77%)	Type of housing - flat (8.41%)
Feature distribution	Multiple features	Location and type
Market stability	High	Moderate

V. Web application

An interactive web application was developed to demonstrate practical deployment of the trained models. Main application features are listed below.

Dashboard view

- Real-time model performance metrics for both markets;

- Visual comparisons of R^2, RMSE, and MAE across models;

- Cross-market performance analysis;

- Feature importance visualizations.

Price calculator

- Ames Calculator: users input Overall Quality (1-10), Living Area (sqft), Year Built, Garage Capacity, Basement Area, and Bathrooms. The system applies the XGBoost model to generate price estimates with confidence intervals based on ±RMSE ($18,632).

- Malaysia Calculator: users input Property Size (sqft), Location (premium areas like KLCC, Bukit Bintang), Rooms, Bathrooms, and Parking Bays. The XGBoost model provides estimates with ±RM 156,786 confidence ranges.

Technical implementation

- Single-page application using HTML5, CSS3, and vanilla JavaScript;

- Responsive design supporting desktop and mobile devices;

- Client-side calculation using simplified model approximations based on feature importance weights;

- Real-time result visualization with confidence interval displays.

Access: the web application is hosted in the project repository at https://github.com/ZoNNeRon/Real-Estate-Price-Prediction/

VI. Discussion

Model performance insights

The 48-point performance gap between markets (Ames: 90.67% vs Malaysia: 42.71% R^2) reveals the critical impact of data quality and feature engineering on prediction accuracy. Ames dataset's comprehensive feature set and consistent collection methodology enable models to capture intricate pricing patterns. Malaysia dataset's simpler feature space and potential data collection inconsistencies limit model effectiveness.

XGBoost's consistent superiority across both markets (6-11 percentage points above alternatives) demonstrates the algorithm's robust capability to capture non-linear relationships and interaction effects. The gradient boosting mechanism effectively handles both distributed feature importance (Ames) and concentrated patterns (Malaysia).

Feature importance patterns

Ames market exhibits distributed feature importance with the top feature accounting for 32.77%, indicating multifactorial pricing where quality, size, amenities, and location collectively determine value. This aligns with mature market characteristics where buyers consider comprehensive property attributes.

Malaysia market shows concentrated importance, suggesting location-driven pricing typical of developing markets where geographic premium dominates other factors. The prominence of Selangor and Kedah reflects established area perception.

Limitations

1. Temporal Validity: Ames data (2006-2010) may not reflect current market conditions; Malaysia data (2025) requires validation.

2. Geographic Scope: results may not generalize to other markets without retraining.

3. Feature Availability: deployment requires consistent feature collection matching training data.

4. Model Approximation: web application uses simplified calculations rather than full models for performance.

VII. Conclusion

This research demonstrates the effectiveness of machine learning, particularly XGBoost, for real estate price prediction across diverse markets. Key findings include:

1. XGBoost achieves superior performance in both mature ($R^2 = 0.9067$) and emerging ($R^2 = 0.4271$) markets

2. Data quality and feature engineering critically impact model effectiveness (48-point R^2 difference)

3. Feature importance patterns reflect market maturity: distributed (Ames) vs. concentrated location-driven (Malaysia)

4. Interactive web applications enable practical deployment and end-user accessibility

Future work should explore:
- Deep learning architectures for automatic feature learning;
- Transfer learning techniques for cross-market model adaptation;
- Time-series analysis for price trend prediction;
- Integration of external economic indicators and market sentiment data.

The developed web application demonstrates the feasibility of deploying machine learning models for real-world real estate applications, providing accessible tools for property valuation and investment decision support.

References

[1] Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications*, 38(4), 3383-3386.

[2] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934.

[3] Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

[4] Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772-1778.

[5] Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 6-10.

[6] Phan, T. D. (2018). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. *Proceedings of the 2018 International Conference on Machine Learning and Data Engineering*, 8-12.