

A Comparative Analysis of Machine Learning Models for the Klang Valley Shop Price Index (KV-SPI)

Author: Sr Fung Cheng Loong, Deputy Director, Property Index Division, National Property Information Centre (NAPIC) can be contacted via email cheng@jpph.gov.my. This article was originally published by Sr Dr. Junainah binti Mohamad et al., 2025 entitled “Constructing a Malaysian Shop Price Index (MSPI) Using Machine Learning Analytics Method” funded by the National Real Property Research Coordinator (NAPREC), National Institute of Valuation (INSPEN), Valuation and Property Services Department (JPPH), Ministry of Finance, Malaysia.

1.0 Introduction

The Klang Valley, encompassing Kuala Lumpur and Selangor, stands as Malaysia's primary economic and commercial hub. The development of a reliable and accurate price index for commercial properties within this region is of strategic importance for investors, policymakers, and market analysts. Such an index serves as a critical barometer for market health, enabling informed decision-making and risk assessment. This report details a rigorous evaluation of modern machine learning methodologies against conventional statistical approaches to construct a robust shop price index.

The scope of this analysis is precisely defined to ensure data integrity and relevance. The geographical focus is the **Klang Valley**, the property type is limited to **1-4 storey terraced shops**, and the analysis period covers transaction data from **2013 to 2024**. This focused scope, determined by an expert panel, ensures a consistent and comparable dataset for modeling.

The primary objective of this report is to systematically evaluate the performance of three machine learning models—**Linear Regression, Random Forest, and XGBoost**—and compare their predictive power against conventional hedonic indexing methods. The evaluation is multifaceted, incorporating quantitative statistical metrics and qualitative visual diagnostics to provide a comprehensive performance assessment.

This report will culminate in a data-driven recommendation for the most accurate and robust model. The chosen model will form the basis for the official **Klang Valley Shop Price Index (KV-SPI)**, providing a new standard for tracking commercial property price movements in the region.

2.0 Methodology and Data Framework

This section outlines the foundational elements of the study, detailing the data sources, the different analytical approaches considered, and the specific protocols used for model development and evaluation. A rigorous and transparent methodology is crucial for establishing the credibility and reliability of the resulting price index.

2.1 Data Sourcing, Scope, and Preparation

The foundation of this analysis is a rich dataset of property transactions sourced from **VIS (Valuation Information System)**. Following an expert panel review, the data selection was refined to meet specific criteria designed to ensure consistency and minimize statistical noise:

- **Geographical Scope:** Restricted to the Klang Valley (Kuala Lumpur and Selangor).
- **Property Type:** Focused exclusively on terraced shops.
- **Property Specifications:** Included only properties with 1 to 4 storeys.

From the comprehensive transaction data, a wide array of variables was selected for the index development. This comprehensive feature set is essential for the hedonic model to control for the wide variety of property characteristics and isolate the pure effect of time on price. The variables used include:

- YEAR, QUARTER, STATE, REGION, MUKIM, SCHEME, LAND AREA, MFA, AFA, POSITION OF SHOP, CONDITION OF SHOP, TYPE OF SHOP CONSTRUCTION, AGE OF SHOP, TERRAIN, STREET LAYER, CLASSIFICATION OF LOCATION, CONDITION OF LOCATION, LAND USE, DISTANCE TO NEAREST CITY, TENURE, PURCHASER STATUS, FIRST TRANSFER, TYPE OF LOT, TYPE OF OWNERSHIP, NEW AREA, PRICE, PRICE LOG, PRICE PSM, PRICE PSM LOG

2.2 Index Construction Methodologies: A Comparative Overview

Two distinct classes of methodologies were evaluated for constructing the price index.

2.2.1 Conventional Hedonic Price Index Methods

These models represent the traditional approach to property price indexing and are commonly used by organizations like the Parliamentary Budget Office - Real Estate Index (PBO-RI) and the Malaysian House Price Index (MHPI). The core principle involves using regression coefficients to account for property characteristics. The three conventional methods considered are:

- **Conventional Hedonic Price Index:** Utilizes regression coefficients from a single hedonic model where characteristics are held constant over time.
- **Chained Time-Varying Parameter Hedonic Index:** Estimates hedonic models over short, successive periods, with indices "chained" together to track sequential price changes.
- **Laspeyres Time-Varying Parameter Hedonic Index:** Integrates base-period weights with time-varying coefficients to construct the index.

2.2.2 Non-Parametric Machine Learning-Based Methods

This modern approach leverages powerful, non-parametric algorithms to model complex relationships within the data. Instead of relying on regression coefficients, these methods derive the price index directly from **predicted price changes**. The primary models in this category are:

- **Random Forest:** An ensemble method that builds a multitude of decision trees and aggregates their predictions, enhancing accuracy and robustness.
- **Gradient Boosting (XGBoost):** An advanced ensemble technique that builds trees sequentially, with each new tree correcting the errors of the previous ones, resulting in highly accurate predictions.

2.3 Model Development and Evaluation Protocol

A structured, nine-stage workflow was implemented, guiding the process from initial data acquisition to the final index reporting.

KV-SPI Development Stages:

1. Data Collection
2. Data Cleaning & Preparation
3. Exploratory Analysis
4. Model Development (Machine Learning and Benchmark)
5. Prediction of Base Year Values (2018)
6. Prediction of Yearly/Quarterly Values (2013–2024)
7. Model Validation
8. Computation of the Index
9. Aggregation & Reporting of KV-SPI

For the machine learning models, a standardized technical setup was used to ensure fair comparison and reproducibility. The dataset was split into training and testing sets using a **70/30 ratio**, with a random_state of 42 applied to ensure the split remains consistent across all model runs.

An initial evaluation was conducted on a broad set of machine learning algorithms, including Support Vector Regressor (SVR), Decision Tree, and a Multi-Layer Perceptron (MLP). Based on superior preliminary performance, the following three models were selected for this in-depth analysis:

1. **Linear Regression**
2. **Random Forest**
3. **XGBoost**

The KV-SPI was calculated using the predictions from the best-performing model based on the following formula, with **2018 established as the base year (Index = 100)**:

$$KVSPIt = \frac{\text{Predicted Price}_t}{\text{Predicted Price}_{2018}} \times 100$$

This detailed framework ensures that the subsequent analytical results are both robust and directly comparable.

3.0 Quantitative and Qualitative Model Performance

This section presents the core evaluation of the machine learning models. The assessment combines hard statistical metrics with visual diagnostics to provide a multi-faceted view of each model's predictive power and its ability to capture the complex dynamics of the property market.

3.1 Statistical Performance Metrics

To objectively measure predictive accuracy, two key metrics were calculated on the test dataset: Root Mean Squared Error (RMSE) and R-squared (R^2). RMSE measures the average magnitude of the prediction errors, while R^2 indicates the proportion of variance in the actual prices that is predictable from the features.

Model	RMSE (Test)	R^2 (Test)
Linear Regression	0.20	0.9247
Random Forest	0.01	0.9999
XGBoost	0.02	0.9988

The results clearly demonstrate the superior performance of the ensemble-based models. **Random Forest** and **XGBoost** achieved near-perfect R^2 scores of 0.9999 and 0.9988, respectively, indicating they can explain over 99.8% of the variability in shop prices. Their exceptionally low RMSE values (0.01 and 0.02) signify minimal prediction error. In stark contrast, **Linear Regression**, while still performing reasonably well with an R^2 of 0.9247, exhibited a significantly higher RMSE of 0.20, highlighting its limitations in accurately modeling the market's complexities.

3.2 Visual Diagnostics of Predictive Accuracy

Beyond statistical scores, visual inspection provides intuitive insights into model behavior and fit.

3.2.1 Predicted vs. Actual Price Distributions

Histograms comparing the distribution of predicted prices (orange) against actual prices (blue) reveal how well each model captures the underlying data structure.

- **Linear Regression:** The predicted distribution for Linear Regression shows a noticeable deviation from the actuals, particularly at the mode of the distribution. It fails to precisely capture the peak frequency and demonstrates a slight distributional shift, which is characteristic of underfitting in a model constrained by linear assumptions.
- **Random Forest & XGBoost:** Both models exhibit a near-perfect overlap between the predicted and actual distributions. This demonstrates their advanced capability to model non-linear relationships and accurately replicate the true price distribution within the Klang Valley shop market.

3.2.2 Predicted vs. Actual Price Scatter Plots

Scatter plots of predicted versus actual prices offer a direct visual assessment of prediction accuracy for individual transactions.

- **Linear Regression:** The plot shows a noticeable dispersion of data points around the ideal diagonal trend line. This spread indicates a higher level of prediction error for many individual properties.
- **Random Forest & XGBoost:** For both models, the data points form a tight, clean line along the diagonal. This exceptional alignment signifies that the predicted prices are almost identical to the actual prices, reinforcing the models' outstanding accuracy.

The combined quantitative and qualitative evidence overwhelmingly confirms the superior predictive capabilities of Random Forest and XGBoost, making them prime candidates for constructing the final price index.

4.0 Price Index Validation: Aggregate and Regional Trends

A reliable price index must not only be statistically accurate but also reflect real-world market trends. This section validates the generated indices against actual price movements at both the macro level (entire Klang Valley) and the micro level (specific administrative regions) to ensure the models are robust and practical for real-world application.

4.1 Klang Valley Aggregate Index Performance

The aggregate index chart provides a high-level comparison of all tested methodologies against the actual price trend for the entire Klang Valley.

This visualization reveals critical performance differences:

- **Conventional Methods:** The **Laspeyres Hedonic** index demonstrates extreme volatility and unreliability, highlighted by an anomalous and unrealistic spike to an index value of **162.0 in 2022**. This erratic behavior makes it unsuitable for practical use.
- **Machine Learning Models:** In sharp contrast, the indices generated by **Random Forest** and **XGBoost** track the actual price index line almost perfectly, capturing every inflection point with high fidelity. The **Linear Regression** model follows the general trend but shows consistent, albeit minor, deviations from the actual index line.

4.2 Regional Index Performance Validation

To test the models' robustness, their performance was validated across various distinct sub-markets within the Klang Valley. The charts for key regions like Kuala Lumpur North and Petaling are representative of the findings across the entire dataset.

The regional analysis consistently reinforces the findings from the aggregate level:

- Across every region—from Kuala Lumpur Central and Petaling to Gombak and Klang—the **Random Forest** and **XGBoost** models demonstrate a superior

ability to trace the nuanced price movements and turning points unique to each local market.

- For example, in **Kuala Lumpur North**, both Random Forest and XGBoost accurately captured the sharp market peak with an index value of approximately 160 in 2022 and the subsequent correction, mirroring the actual trend precisely.
- Similarly, in the **Petaling** sub-market, both models successfully captured the 2021 peak (index 112.8), the subsequent dip in 2022 to 105.6, and the sharp recovery to a new high of over 113 in 2023, while the Linear Regression model consistently lagged behind these turning points.

The overwhelming visual evidence from both aggregate and regional validations confirms that the advanced machine learning models provide a far more accurate and reliable representation of property price dynamics.

5.0 Synthesis: Why Machine Learning Excels for Price Index Construction

The superior performance of Random Forest and XGBoost is not arbitrary; it is a direct consequence of their fundamental architectural differences from conventional linear models. These non-parametric, ensemble-based methods are inherently designed to model high-dimensional, non-linear relationships without being constrained by the rigid statistical assumptions that limit traditional hedonic techniques.

Modeling Attribute	Random Forest / XGBoost Approach	Conventional (Laspeyres/Linear) Approach
Non-linearity	Captures complex non-linear relationships and feature interactions automatically, reflecting true market behavior.	Assumes linear or log-linear relationships, limiting its ability to model complex interactions.
Robustness to Outliers	High robustness through ensemble averaging (Random Forest) and gradient-based optimization (XGBoost).	Highly susceptible to influence from extreme values or influential data points, which can skew results.
Feature Selection	Performs automatic feature selection and provides insights into variable importance, improving model efficiency.	Requires manual feature selection and offers no systematic assessment of feature influence.
Handling High-Dimensional Data	Effectively manages large datasets with many, often correlated, variables without issues like multicollinearity.	Often limited by problems of multicollinearity and dimensionality, requiring more data preprocessing.
Predictive Accuracy	Achieves significantly higher predictive accuracy through ensemble learning and adaptive optimization techniques.	Accuracy is often constrained by model specification and violations of underlying statistical assumptions.
Weighting and Updating	Adaptively updates weights and decision structures based on	Uses static, base-period weights, making it less responsive to evolving market dynamics.

	training data and residual errors, learning from the data.	
Data Assumptions	Makes few to no assumptions about data distribution (e.g., normality) or variance (e.g., homoscedasticity).	Relies on strict assumptions of normality, independence, and constant variance, which are often not met in practice.

This comparative analysis makes it clear that the technical superiority of the machine learning approach is not incidental but is rooted in its fundamental design, making it inherently better suited for the complex task of property price index construction.

6.0 Conclusion and Final Recommendation

This report has conducted a comprehensive evaluation of multiple methodologies for constructing a shop price index for the Klang Valley. The findings are conclusive. Both quantitative metrics, including R-squared values exceeding 0.998 and minimal Root Mean Squared Error, and extensive visual validation at the aggregate and regional levels demonstrate the profound superiority of the **Random Forest** and **XGBoost** models. These advanced techniques vastly outperform both the simpler Linear Regression model and conventional hedonic methods, which proved to be either less accurate or unacceptably volatile.

Based on the comprehensive evidence presented, the final recommendation is unequivocal.

Recommendation:

It is formally recommended that either the **Random Forest** or the **XGBoost** model be adopted for the official construction and annual publication of the **Klang Valley Shop Price Index (KV-SPI)**.

This recommendation is justified by the models' unparalleled predictive accuracy, their proven robustness in tracking diverse and dynamic regional market trends, and their intrinsic ability to model the complex, non-linear nature of property markets. Adopting this machine learning-based approach will provide stakeholders with a significantly more reliable, accurate, and nuanced tool for understanding the Klang Valley's commercial property market.