Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 7(1) 2026, 22 – 36.

Mathematical Sciences and Informatics Journal

Real Estate Analytics and Price Prediction Using Automated Machine Learning

Muhammad Syazani Mohamad Jaffar[1], Noor Latiffah Adam[1*], Sharifalillah Nordin[1]

[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor*

ARTICLE INFO

Article history:
Received 14 October 2025
Revised 5 January 2026
Accepted 28 January 2026
Online first
Published 30 April 2026

Keywords:
Automated Machine Learning
AutoML
PyCaret
Real Estate Analytics
Real Estate Price Prediction

DOI:
10.24191/mij.v7i1.11568

ABSTRACT

Navigating the Malaysian real estate market is increasingly complex due to the dynamic interplay of localised socioeconomic factors and high-dimensional variables. Addressing the demand for objective decision-support tools, this research develops a comprehensive analytics and price prediction framework tailored for four primary economic hubs: Selangor, Kuala Lumpur, Penang, and Johor. Diverging from conventional manual modelling, the methodology employs an Automated Machine Learning (Auto ML) approach via the PyCaret library to systematically explore feature spaces while minimising human bias during algorithm selection. Evaluation of various regression architectures utilised standard benchmarks, specifically Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination R^2. Empirical results indicate that ensemble-based models consistently exceed the performance of alternative approaches; Random Forest achieved the highest accuracy for Selangor, Penang, and Johor, while the Extra Trees Regressor emerged as the superior model for Kuala Lumpur. Furthermore, a unified multi-regional model demonstrated exceptional generalisation capabilities, recording an R^2 of 0.893, an RMSE of RM68,497.17, and an MAE of RM46,104.28 on a combined dataset. The primary contribution of this work lies in the seamless architectural integration of these high-performance models into a real-time, web-based dashboard, transitioning the research from static academic analysis to a dynamic, user-facing system. Ultimately, these findings confirm that an Auto ML-driven framework simplifies complex analytical workflows and delivers the precision required to support data-driven decision-making and transparent property valuation in a developing economy.

[1*] Corresponding author. *E-mail address*: *latiffah508@uitm.edu.my*
https://doi.org/10.24191/mij.v7i1.11568

1. INTRODUCTION

Property valuation is currently moving through a seismic shift, trading manual, intuition-heavy estimations for the precision of automated intelligence. While traditional methods are typically anchored in cost-based or comparative logic. It often stumbles when faced with market volatility or subjective bias. Machine Learning (ML) offers a much-needed objective alternative. As Ja'afar et al. (2021) point out, these computational frameworks do more than just process numbers; they distil actionable insights from massive, historical datasets to forecast trends with a level of accuracy that was previously unreachable.

This precision is not just a luxury but is a fundamental requirement for economic stability. Real estate serves as the bedrock of both individual wealth and national fiscal health, which means that everyone from the private first-time buyer to the high-level urban planner needs a reliable risk-mitigation tool. The pace of modern urbanisation has introduced fluctuations so rapid that traditional forecasting simply cannot keep up with (Wu, 2024). This is the reason that adopting advanced ML is no longer a niche academic pursuit, but is a critical necessity for transparent market policymaking (Li, 2024). Recent breakthroughs have already proven that algorithms like Support Vector Machines (SVM) and Light GBM consistently outperform old-school linear models (Tandon, 2024). Now, the rise of Automated Machine Learning (Auto ML) is taking this a step further. By utilising Python-based libraries like PyCaret, we can turn data cleaning, feature engineering, and hyperparameter tuning into a single, seamless workflow. This "low code" approach is not just about speed; it is about increasing experimental throughput while maintaining the rigorous standards required for property analytics (Mantaw & M, 2024).

However, applying these global tools to the Malaysian context presents its own set of so-called "homegrown" challenges. Our market is currently navigating a persistent "overhang" of unsold units and significant information gaps, especially within major cities in Johor and Selangor (NAPIC, 2025). A major roadblock is the temporal lag in data reporting, which often happens by the time transaction records go public. At that time, the market sentiment had already shifted. Moreover, existing Malaysian research tends to be overly focused on Kuala Lumpur, ignoring the unique socio-economic DNA of other regions (Mohd, 2019). Most of these models also suffer from being "black box", where they give a number but offer no transparency to the actual human stakeholders who need to understand *why* a price is what it is.

This study steps into that gap. By creating an ML framework that specifically works across different regions and is specially designed for Malaysia's main economic area such as Selangor, Kuala Lumpur, Penang, and Johor. Here, we want to offer a more detailed and accurate tool for the industry. These four regions are the heartbeat of the national market; Selangor alone accounted for 22% of all Malaysian transactions in 2023 (Valuation and Property Services Department, 2024). By training our models on data collected over time from January 2021 to June 2024, we ensure the system captures the post-pandemic recovery phase, offering a modern and detailed view of the property values that wider, national-level studies usually overlook.

2. RELATED WORK

Rapid expansion within the Malaysian property sector has intensified the demand for precise valuation frameworks, which serve as the foundation for strategic decision-making among developers and institutional investors. Reliable forecasting models provide the necessary foresight to navigate shifting market conditions, offering stakeholders a quantitative basis for risk assessment (Ja'afar et al., 2021). Traditional statistical approaches often fail to capture the intricacies of modern property markets; consequently, Machine Learning (ML) has emerged as the premier analytical standard due to its proficiency in identifying non-linear patterns within heterogeneous datasets. These algorithms utilize advanced pattern recognition and statistical inference to transform raw market data into actionable predictive models (Li, 2024).

https://doi.org/10.24191/mij.v7i1.11568

Integrating sophisticated analytical architectures remains a fundamental requirement for modern real estate research. Mathotaarachchi et al. (2024) emphasises that adopting ML methodologies significantly elevates the precision of future price estimates, facilitating more robust investment strategies and optimised pricing structures. Recent innovations in Automated Machine Learning (Auto ML) further democratize access to these tools. Frameworks like PyCaret enable researchers with limited programming backgrounds to deploy high-level analytical systems with minimal friction (Quispe et al., 2024). By consolidating data cleaning, algorithm comparison, and performance evaluation into a streamlined workflow, the PyCaret library accelerates the transition from raw data to model deployment without sacrificing computational rigor.

2.1 Real Estate Price Prediction

Predicting property values remains a primary concern for a diverse range of market participants, including institutional investors, developers, and residential buyers. The persistent growth and inherent stability of the property sector heighten the sensitivity of these forecasts (Wu, 2024). Traditional approaches to this problem, as noted by Phan (2018), involve synthesising various attributes such as physical dimensions, geographic coordinates, onsite amenities, and macroeconomic indicators. Historically, these estimates relied on statistical frameworks that assumed linear correlations between a property's price and its defining characteristics.

However, the increasing complexity and dynamic nature of modern property markets necessitate the adoption of more robust valuation architectures. Industry stakeholders—ranging from policymakers to homeowners—depend on these predictions to facilitate high-stakes decisions (Mathotaarachchi et al., 2024). Investors utilise predictive tools to quantify potential returns, while developers rely on them to assess the viability of new projects, and buyers use them to benchmark acquisition costs. Regional economic shifts, localised policy changes, and global financial trends introduce significant volatility into the market, complicating the quest for accuracy. Consequently, the methodologies used to navigate these challenges have transitioned away from rudimentary statistical models toward the sophisticated capabilities of artificial intelligence and machine learning.

2.2 Machine Learning

Implementing Machine Learning (ML) allows computational systems to derive insights from data and construct predictive frameworks without the need for explicitly hard-coded logic. These algorithms excel at identifying intricate, non-linear correlations within vast datasets, a capability that is particularly advantageous in the real estate sector, where property values are influenced by a multitude of intersecting variables (Li, 2024). While ML encompasses various paradigms, including supervised and unsupervised learning (Ja'afar et al., 2021), supervised learning remains the most effective approach for property valuation. By training algorithms on structured regression datasets with known target outcomes, these models learn to forecast future prices based on established historical patterns.

Specific techniques such as Decision Trees, Random Forests, and XGBoost have gained significant traction in the literature due to their robust stability and predictive precision. Empirical studies conducted by Li (2024), Mohd et al. (2019), and Zulkifley (2020) confirm that these ensemble-based methods possess a unique aptitude for processing the non-linear dynamics of housing markets. Furthermore, these architectures can simultaneously evaluate multiple independent variables, providing a comprehensive view of the diverse factors governing real estate appraisal.

Selecting the optimal configuration for these models, however, remains a resource-heavy endeavour. Identifying the ideal hyperparameter settings for a given algorithm typically requires a rigorous, iterative cycle of training and evaluation. Because this process often relies on trial-and-error or heuristic strategies, the search space expands exponentially as more candidate models are introduced. Consequently,

https://doi.org/10.24191/mij.v7i1.11568

pinpointing the highest-performing model within fixed time and computational limits presents a significant technical hurdle for researchers.

2.3 Automated Machine Learning

Streamlining the development of predictive frameworks is a primary advantage of PyCaret, an open-source Automated Machine Learning (Auto ML) library. This Python-based platform serves both novice users and experienced practitioners by accelerating the transition from raw data to functional models. Mantaw (2024) demonstrates that employing PyCaret significantly hastens the selection of optimal features, algorithms, and hyperparameter configurations for price forecasting. Implementing Auto ML frameworks effectively reduces the time required to establish high-performance models, thereby enhancing the overall design and execution of ML experiments (Annamalai et al., 2023).

Operating as a low-code environment, this library supports comprehensive, end-to-end ML workflows (Huynh et al., 2023). Such automation is particularly effective for regression and time-series tasks, which are fundamental to accurate property price estimation. Research by Quispe et al. (2024) highlights that PyCaret facilitates the simultaneous comparison of up to 27 distinct ML algorithms. Its architectural versatility encompasses a wide array of supervised and unsupervised techniques; for instance, the regression suite includes everything from Linear and Ridge models to sophisticated tree-based ensembles and boosting variants. Consequently, the library simplifies the deployment phase, making complex analytical tools more accessible for real-world applications.

Existing Malaysian property studies (Mohd et al., 2019) have already validated the utility of Machine Learning, yet many continue to rely on manual, labour-intensive workflows. These traditional methods are often hampered by "algorithm bias," where researchers pre-emptively select specific models like Neural Networks or Support Vector Machines without conducting exhaustive cross-algorithm benchmarking. This study diverges from such restricted, single-region methodologies by adopting an Auto ML framework. This shift is essential for objectively navigating the high-dimensional variables of the Malaysian market and reducing human error during the tuning process. Furthermore, this research addresses the operationalisation gap found in previous academic-only frameworks by integrating ensemble models into a web-based dashboard. Validating model performance across heterogeneous regions like Johor and Penang ensures that the resulting system is not only accurate but also scalable for practical decision support.

3. MATERIALS AND METHODS

The real estate price predictions are based on historical transaction data for Selangor, Kuala Lumpur, Penang, and Johor from January 2021 to June 2024, which consists of data regarding property type, district, ‘mukim’, area, date of transaction, land area, tenure, and transaction price (NAPIC, 2025).

3.1 Data Processing and Feature Engineering

Preparing the raw data for analysis involved a multi-stage pre-processing pipeline designed to maximise consistency and model compatibility. Maintaining dataset integrity required the use of mean substitution for imputing missing values, a technique chosen to prevent the loss of critical observations. Enhancing the predictive power of the model involved a dedicated feature engineering phase, where raw data was transformed into more descriptive variables; for example, a standardised metric for price per square foot was calculated to normalise valuation across different property sizes. Scalability and convergence issues were further addressed by normalising all continuous features. Reducing these scale disparities ensures that the algorithms treat each variable with appropriate weight, ultimately leading to more stable and accurate performance during the training phase. The dataset consists of 255,184 records. Fig. 1 shows a sample of the dataset used for the project.

https://doi.org/10.24191/mij.v7i1.11568

	A	B	C	D	E	F	G	H	I	J	K
1	Property Type	District	Mukim	Scheme Name/Area	Transaction Date	Tenure	Land/Parcel Area	Main Floor Area	Unit Level	Transaction Price	
2	Detached	Hulu Selangor	Bandar Hulu Yam (Baru)	KAMPONG BARU ULU YAM	1/10/2022	Leasehold	352	22		100000	
3	1 - 1 1/2 Storey Terraced	Kuala Lumpur	Mukim Setapak	TAMAN WARISAN PELANGI	1/6/2022	Freehold	24.21	24		183000	
4	1 - 1 1/2 Storey Terraced	Kuala Selangor	Batang Berjuntai	PERUMAHAN BERJUNTAI TIN	1/10/2022	Leasehold	162.58	37		150000	
5	1 - 1 1/2 Storey Terraced	Kuala Selangor	Batang Berjuntai	PERUMAHAN BERJUNTAI TIN	1/12/2022	Leasehold	162.58	37		123000	
6	Detached	Sabak Bernam	Pekan Sungai Besar	PEKAN SUNGAI BESAR	1/12/2023	Leasehold	750	37		220000	
7	Low-Cost House	Kuala Selangor	Batang Berjuntai	PERUMAHAN BERJUNTAI TIN	1/12/2023	Leasehold	162.58	37		170000	
8	Low-Cost House	Kuala Selangor	Batang Berjuntai	PERUMAHAN BERJUNTAI TIN	1/10/2022	Leasehold	182.08	37		130000	
9	Low-Cost House	Kuala Selangor	Batang Berjuntai	PERUMAHAN BERJUNTAI TIN	1/2/2023	Leasehold	162.58	37		120000	
10	Detached	Gombak	Batu	KAMPONG SELAYANG PANDANG	1/11/2021	Leasehold	371.6	42		300000	
11	Low-Cost House	Klang	Kapar	TMN MERU (R/M MERU III)	1/8/2022	Leasehold	130.06	42		275000	
12	Low-Cost House	Klang	Kapar	TMN MERU (R/M MERU III)	1/10/2022	Leasehold	130.06	42		280000	
13	Low-Cost House	Klang	Kapar	TMN MERU (R/M MERU III)	1/2/2022	Leasehold	130.06	42		200000	
14	Low-Cost House	Hulu Selangor	Bandar Kalumpang	TAMAN MUSABIKA	1/8/2023	Leasehold	111	43		45000	
15	Low-Cost House	Hulu Selangor	Kuala Kalumpang	TAMAN MUSABIKA	1/6/2022	Leasehold	130.06	43		105000	
16	Low-Cost House	Hulu Selangor	Kuala Kalumpang	TAMAN MUSABIKA	1/7/2022	Leasehold	111.48	43		120000	

Fig.1. Sample dataset was used for the project

3.2 System Architecture and Implementation Environment

System architecture relied on a suite of Python-based tools to ensure the efficient development and deployment of the predictive framework. Leveraging PyCaret, an open-source Auto ML library, allowed for the seamless automation of the machine learning pipeline, including data preparation, feature scaling, and the systematic selection of high-performance models. This framework also handled the complex task of hyperparameter optimisation, ensuring each algorithm reached its peak predictive potential with minimal manual intervention. Developing the user-facing portion of the system involved Stream lit, which was utilised to construct a web-based interactive dashboard. This choice enables stakeholders to interact with the models and visualise price forecasts through an intuitive graphical interface. Supporting the entire coding lifecycle, Visual Studio Code (VS Code) served as the primary integrated development environment. This IDE provided the necessary infrastructure for robust coding, iterative testing, and comprehensive debugging, ultimately streamlining the transition from experimental scripts to a functional application.

3.3 Model Training, Evaluation, and Selection

Model training was conducted using the regression module of PyCaret, which automatically compares multiple machine learning algorithms and optimizes their hyperparameters. A variety of regression models were evaluated to identify the most suitable predictive approach for real estate price estimation. PyCaret minimizes manual intervention during preprocessing and model tuning, thereby improving efficiency and reproducibility. Evaluating the predictive capability of the models required partitioning the dataset into distinct training and testing subsets. Robust generalisation to unseen data was ensured through the application of k-fold cross-validation during the initial training phase. Three standard regression metrics served as the primary benchmarks for performance assessment: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Identifying the optimal model involved selecting the algorithm that demonstrated the highest accuracy and lowest error rates on the testing dataset. Once verified, this top-performing model was serialised into a pickle (.pkl) format. This serialisation process facilitated the seamless integration of the trained intelligence into the Stream lit-based dashboard, allowing for real-time price forecasting in a production environment.

3.4 Dashboard Development and Model Deployment

Constructing an interactive web-based dashboard bridged the gap between complex algorithmic output and practical user application. The interface features a streamlined navigation system that allows users to toggle between the homepage, longitudinal market trend analyses, the primary prediction portal,

https://doi.org/10.24191/mij.v7i1.11568

and detailed result summaries. Enhancing the interpretability of the data involved the integration of dynamic visualisations; specifically, bar charts were used to display market shifts, scatter plots to reveal variable relationships, and heatmaps to visualise the underlying feature correlation matrix. Facilitating the prediction process requires a structured input interface where users define property-specific variables such as geographic location, square footage, and building type. Ensuring data integrity within this process involved the implementation of rigorous input validation mechanisms to filter out inconsistent or erroneous entries. Once the user submits the required attributes, the embedded ML engine generates real-time valuation forecasts. In order to move beyond "black-box" predictions, the system supplements each forecast with analytical insights, including feature importance rankings and confidence indicators. These additions provide the transparency necessary for stakeholders to understand the primary drivers behind a specific property valuation.

4. RESULT AND DISCUSSION

Experimental procedures followed two distinct paths to identify the most effective predictive architecture. The primary experiment involved training and testing independent models for each specific region to isolate localised market nuances. Subsequently, the second experiment merged the datasets from all four geographic areas into a unified training set to assess the viability of a consolidated regional model. Maintaining high standards for data integrity required all records to undergo uniform pre-processing before being partitioned. This segmentation utilised a 90% training and 10% testing split, ensuring a substantial volume of data for model learning while reserving a sufficient portion for objective performance validation.

4.1 Experiment 1 Trained and Tested Individual Region Dataset

Results for the Selangor property market indicate that the Random Forest Regressor provided the most accurate price estimations. This model achieved an R^2 of 0.9107, an RMSE of RM78,117.35, and an MAE of RM50,710.33, demonstrating high precision and a strong capacity for generalization across the dataset. The Extra Trees Regressor followed closely, yielding a nearly identical R^2 of 0.9103, though it exhibited slightly higher error margins in both RMSE and MAE.

Both ensemble-based architectures proved highly effective at mapping the intricate, non-linear variables within the region's data. While gradient boosting frameworks such as XGBoost and LightGBM offered faster training durations, they failed to match the accuracy levels achieved by the tree-based ensemble models. Consequently, the Random Forest Regressor stands as the optimal choice for deployment in the Selangor predictive framework (Fig. 2).

Predictive analysis for the Kuala Lumpur region reveals that the Extra Trees Regressor emerged as the superior model. Achieving an R^2 of 0.7933, an RMSE of RM153,497.57, and an MAE of RM91,116.81, this algorithm demonstrated consistent precision. Supporting its reliability are the low RMSLE (0.2916) and MAPE (21.23%) values, which suggest the model maintains a stable performance even across diverse property tiers in the capital.

https://doi.org/10.24191/mij.v7i1.11568

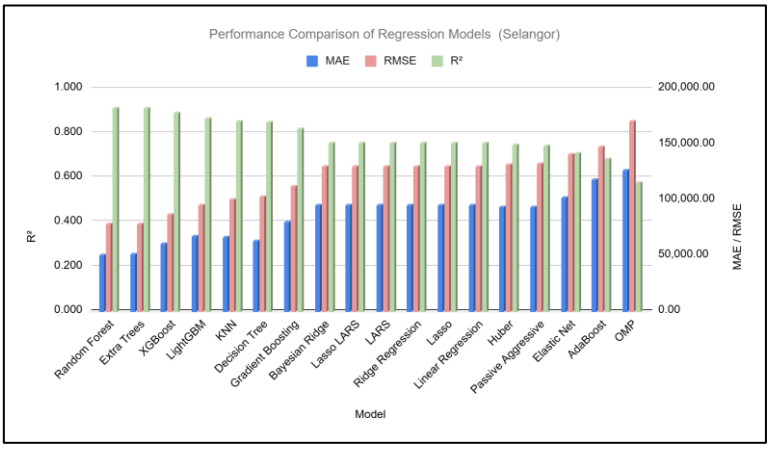

Fig. 2. Experiment 1 House price prediction performance for the Selangor dataset

The Random Forest Regressor occupied the second rank, recording an R^2 of 0.7854, an RMSE of RM156,439.87, and an MAE of RM95,447.88. Both ensemble-based architectures exhibited a significant performance advantage over alternative algorithms tested. Based on this balanced evaluation of accuracy and the ability to generalize to new data, the Extra Trees Regressor serves as the recommended choice for the Kuala Lumpur predictive interface as in Fig. 3.

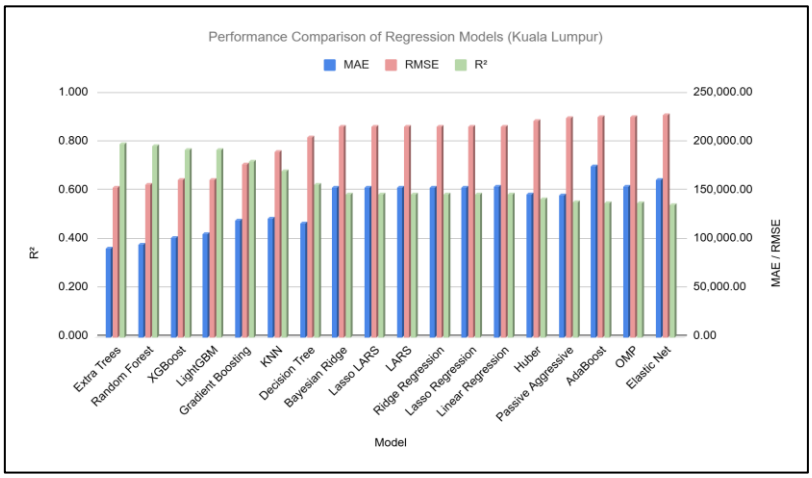

Fig. 3. Experiment 1 House price prediction performance for the Kuala Lumpur dataset

Performance metrics for the Penang property market highlight the Random Forest Regressor as the most effective model. This algorithm achieved an R^2 of 0.9114, an RMSE of RM67,221.47, and an MAE of RM43,770.34, successfully accounting for more than 91% of the total variance in regional housing prices. The Extra Trees Regressor followed with a nearly identical R^2 of 0.9106. While this model exhibited only marginal differences in error metrics compared to Random Forest, it provided the advantage of reduced training duration. Nevertheless, the Random Forest Regressor remains the preferred selection for the Penang dataset, as it offers the most robust equilibrium between high-fidelity predictive accuracy and consistent generalization as in Fig. 4.

https://doi.org/10.24191/mij.v7i1.11568

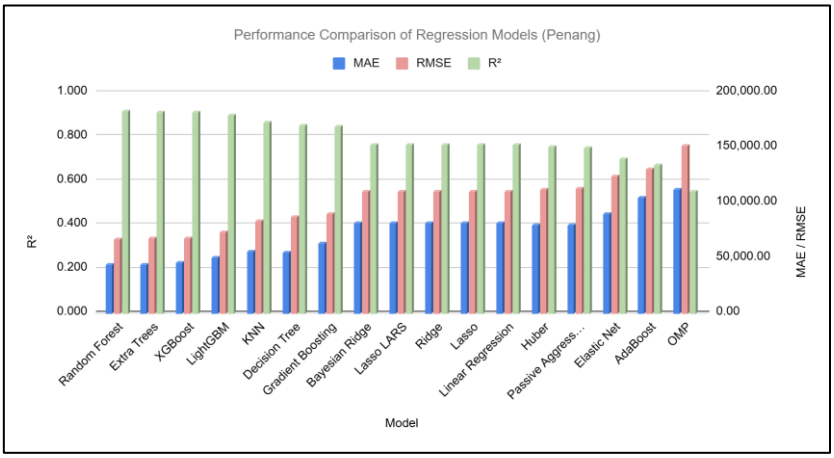

Fig. 4. Experiment 1 House price prediction performance for Penang dataset

Evaluations of the Johor property market identify the Random Forest Regressor as the top-performing architecture. Recording an R^2 of 0.8993, an RMSE of RM68,497.17, and an MAE of RM46,104.28, this model exhibits minimal predictive error and a high degree of reliability when applied to new data. The Extra Trees Regressor reached an R^2 of 0.8967 with an RMSE of RM69,406.96, though it produced a marginally higher MAE in comparison. Both ensemble-based frameworks significantly surpassed the accuracy of alternative algorithms, confirming their suitability for handling the regional complexities of the Johor housing sector. Due to its superior precision and overall robustness, the Random Forest Regressor is the designated model for integration within the Johor analysis module as in Fig. 5.

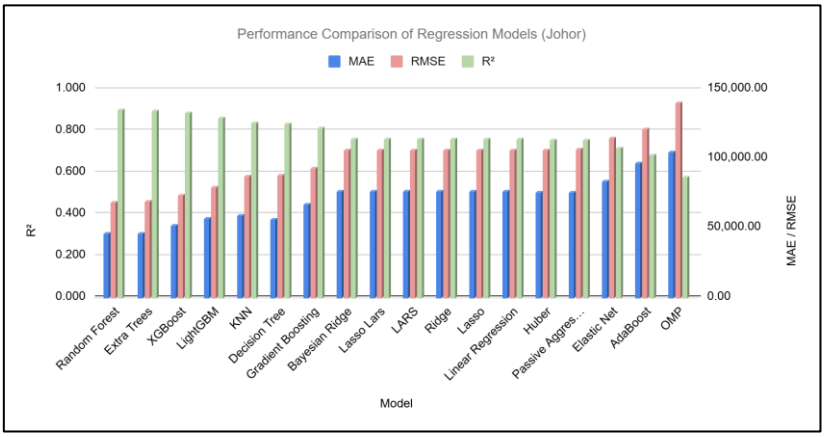

Fig. 5. Experiment 1 House price prediction performance for Johor dataset

4.2 Experiment 2 Trained and Tested Combined Datasets

Analysing the combined data set from all four regions it reveals that the Extra Trees Regressor achieved the most robust results. This model recorded an R^2 of 0.893, an RMSE of RM68,497.17, and an MAE of RM46,104.28, demonstrating a comprehensive ability to represent property price variability while maintaining a low error profile. Following closely, the Random Forest Regressor achieved an R^2 of 0.892, with slightly higher error margins of RM69,406.96 (RMSE) and RM46,306.20 (MAE). XGBoost occupied the third rank (R^2 = 0.851), whereas alternative algorithms like LightGBM, KNN, and Decision

https://doi.org/10.24191/mij.v7i1.11568

Trees exhibited only moderate predictive capabilities. Linear frameworks, including Ridge, Lasso, and standard Linear Regression, performed the least effectively, with R^2 values plateauing at 0.701 or below.

Validating the generalizability of this unified framework involved testing it against an unseen dataset containing records from all four geographic zones. The model maintained an overall R^2 of 0.8979, an RMSE of RM80,441.80, and an MAE of RM50,479.23. Interestingly, the consolidated model's performance in specific regions was nearly identical to, or occasionally surpassed, that of localised models. For instance, in Selangor, the combined model achieved an R^2 of 0.9168 compared to the region-specific score of 0.9181, while in Johor, the RMSE differences were marginal (RM67,777.05 vs. RM66,282.46).

Embracing a single, unified framework yields substantial operational benefits, such as condensed development schedules and minimised maintenance requirements, by removing the necessity for multiple, independently optimised architectures. Following this comprehensive evaluation, the Extra Trees Regressor was selected as the primary computational engine for the integrated property forecasting system as in Fig. 6.

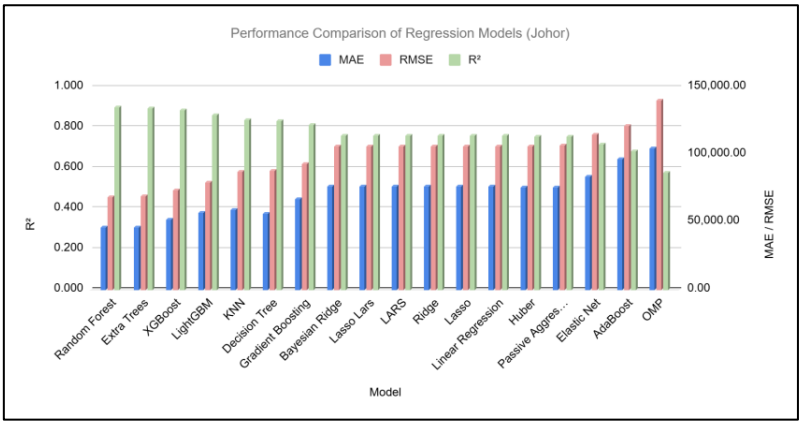

Fig. 6. Experiment 2 House price prediction performance for combined datasets

Detailed comparative analysis in Table 1 further explores the relationship between the multi-regional model and those calibrated for individual states. Although localised versions demonstrate a slightly superior fit for their respective datasets, the integrated model delivers a reliable level of predictive precision throughout the entire study area. This high degree of generalizability reinforces its suitability for large-scale implementation across diverse market environments. Adopting a single, integrated framework offers several strategic advantages, including a streamlined development lifecycle and reduced computational overhead. This approach eliminates the necessity for repetitive fine-tuning of multiple independent architectures, thereby lowering both initial development time and long-term maintenance requirements. Based on this balance of strong performance on novel data and high operational efficiency, the combined regional model was designated as the core engine for the final real estate prediction system.

https://doi.org/10.24191/mij.v7i1.11568

Table 1. Comparison of house price prediction performance with combined datasets

Region	Model	MAE	RMSE	R^2
Selangor	Combined Region Model	49,238.61	74,362.28	0.9181
	Selangor Model	49,813.89	74,942.67	0.9168
Kuala Lumpur	Combined Region Model	88,359.60	148,677.16	0.7943
	KL Model	88,719.87	146,662.44	0.7998
Penang	Combined Region Model	41,783.78	64,457.28	0.9144
	Penang Model	42,414.56	65,975.83	0.9103
Johor	Combined Region Model	45,718.94	67,777.05	0.9025
	Johor Model	44,990.38	66,282.46	0.9068

4.3 Prediction Results and Visual Evaluation

Evaluating the practical reliability of the proposed prediction system involves a visual assessment of the Extra Trees Regressor, specifically focusing on its predictive consistency and error distribution. In Fig. 7 displays a scatter plot comparing actual transaction prices against the model's forecasted values. Although minor deviations appear within the higher price brackets, such variances are common in real estate modelling and typically stem from luxury assets or unique property features that occur less frequently in the training data. The plot overall demonstrates a high level of alignment between observed and predicted prices, confirming the model's robust capacity for generalization.

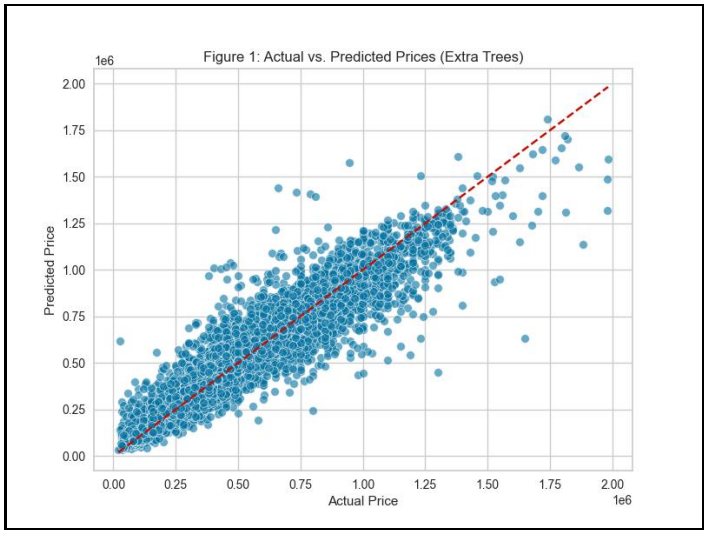

Fig. 7. Scatter plot of actual vs predicted property prices based on Extra Tree model

https://doi.org/10.24191/mij.v7i1.11568

Examining model adequacy further requires a detailed residual analysis. A correctly fitted regression model typically produces residuals that are randomly dispersed around a zero-mean horizontal axis, indicating that the errors are independent and the data structure has been appropriately captured. Fig. 8 illustrates the residual plot for the Extra Trees Regressor, mapping the differences between observed and predicted prices against the forecasted values. The lack of any discernible systematic patterns or "funnelling" effects in the residuals reinforces the stability and reliability of the framework. These collective visual diagnostics confirm that the Extra Trees Regressor serves as a balanced and technically sound architecture for large-scale property valuation.

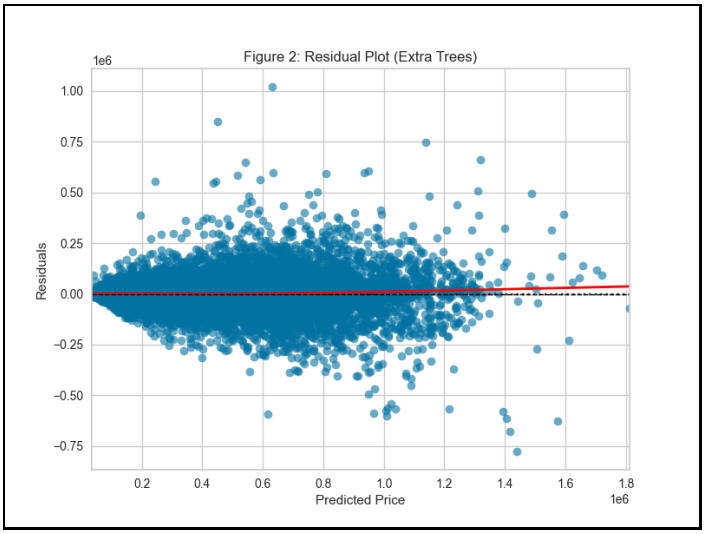

Fig. 8. Residuals Plot for Extra Trees Regressor

4.4 User Interface and System Features

Developing the PropVision Malaysia platform provided a functional interface for the predictive system. Fig. 9 displays the application's homepage, which serves as a central hub outlining the system's core capabilities. This interface emphasises critical property valuation drivers, such as geographic location, land tenure, building type, and total square footage. Market transparency is further enhanced through a dynamic insight card that identifies the state currently recording the highest transaction volume. Featuring this real-time data allows users to immediately understand the prevailing market conditions and identify areas of high property activity across the country.

https://doi.org/10.24191/mij.v7i1.11568

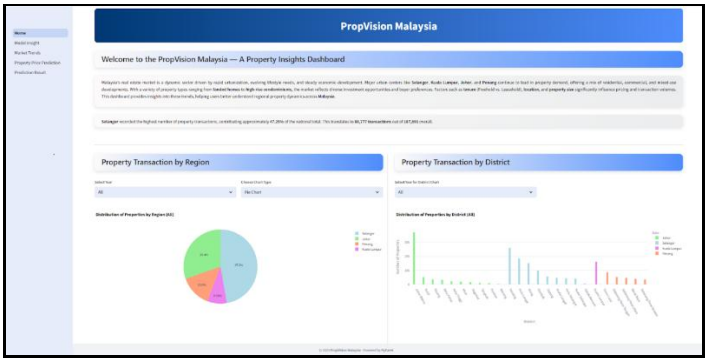

Fig. 9. Home Page of the house price prediction system

Exploring property price dynamics is facilitated through the Market Trends page, which offers an interactive environment for analysing shifts across various timeframes, regions, and building categories. Users possess the ability to refine their analysis by year and state to focus on specific segments of interest.

Visualizing these trends involves a multi-faceted approach: a line chart tracks the evolution of average property values over time, while a corresponding bar chart displays transaction volume per quarter. This combination allows stakeholders to simultaneously evaluate pricing trajectories and overall market liquidity. Furthermore, the interface includes a comprehensive line graph of national average prices alongside a horizontal bar chart illustrating median costs by property type, offering a direct comparison of different market sectors. These combined visualizations empower users to interpret investment patterns and localized market fluctuations within the Malaysian real estate landscape, as illustrated in Fig. 10.

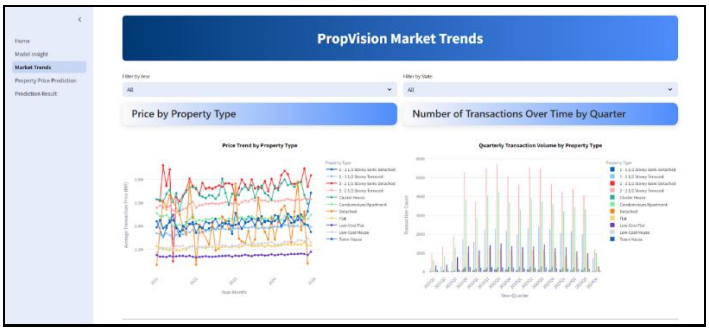

Fig. 10. Market Trends Page

Estimating the market value of specific assets is the primary function of the Property Price Prediction page, as illustrated in Fig. 11. This interface features a structured input form where users define essential variables, including geographic location, building category, land tenure, unit level, and physical dimensions. Ensuring that these estimates align with real-world behaviour involves a backend process that leverages historical transaction data and trained ML architectures. Once the user submits their information, the system executes a multi-step routine consisting of data validation, automated pre-processing, and secure storage before transferring the values to the prediction module. Following this computation, the platform generates a dedicated results page that presents the forecasted price range. Providing this interactive, data-driven environment allows stakeholders to reach informed pricing conclusions based on insights specifically tailored to the unique attributes of their property.

https://doi.org/10.24191/mij.v7i1.11568

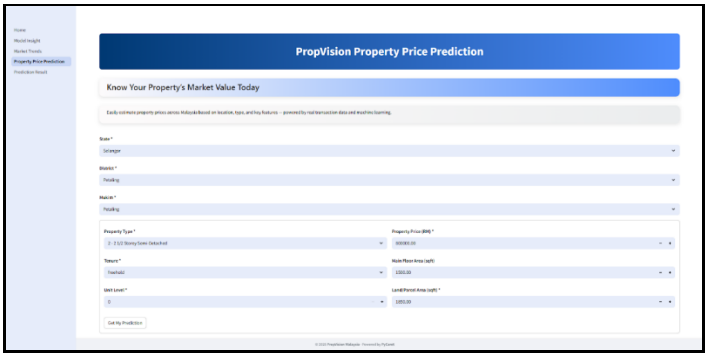

Fig. 11. Property Price Prediction Page

Generating final valuation results is the primary function of the Prediction Result page, as depicted in Fig. 12. This interface displays the estimated market value derived from the specific attributes defined by the user in the preceding step. A pre-trained PyCaret regression model serves as the computational engine, processing the input variables to output a price that reflects contemporary market trends. To ensure transparency, the page features a summary table that explicitly lists the selected property characteristics, enabling immediate verification of the data influencing the forecast. Beyond the numerical forecast, the platform conducts a comparative assessment between the user's expected price and the model-calculated market value. This analysis categorizes the property as reasonably priced, overpriced, or under-priced, utilizing a color-coded visual indicator to emphasize the findings. Implementing this intuitive feedback system facilitates a rapid interpretation of the data, empowering stakeholders to execute timely and well-informed pricing strategies based on objective analytical evidence.

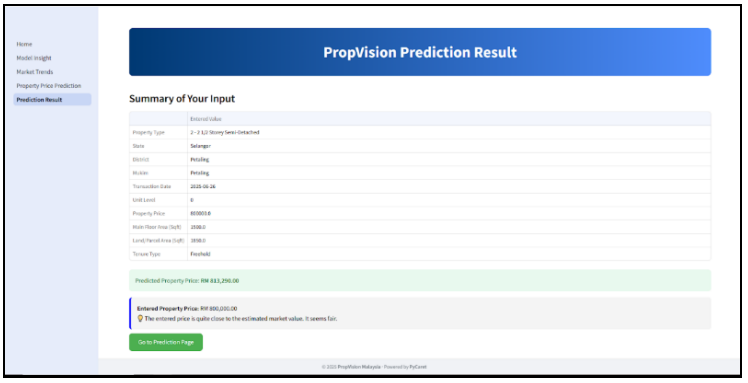

Fig. 12. Prediction Result Page

5. CONCLUSION

This study produces and evaluates a group of ML models aimed at understanding the housing markets in Selangor, Kuala Lumpur, Penang, and Johor. During the experiments, the group of models called ensemble-based architectures, especially Random Forest and Extra Trees, worked the best. They kept performing better than simpler models in terms of accuracy and how steady they were. The data showed that although Random Forest was considered the best method for Selangor, Penang, and Johor, the Extra Trees Regressor took the lead in the more complicated Kuala Lumpur market. One of the most significant findings was the strength of the combined model. By learning from a mix of data, the Extra Trees Regressor achieves an R^2

https://doi.org/10.24191/mij.v7i1.11568

of 0.893, effectively capturing the price fluctuations across all four states with impressive precision. More importantly, this unified model held its own against the region-specific versions, even outperforming them in some instances. This suggests that a centralised approach is not just a quick fix; it is a robust and efficient way to set up a large-scale predictive system while keeping maintenance and computational costs low.

Ultimately, this study proves that Auto ML frameworks like PyCaret are transformative for real estate analytics. By automating the heavy lifting of pre-processing and hyperparameter tuning, we were able to focus on building a system that actually works for the end-user. The resulting dashboard is not just a technical achievement—it is a practical tool that gives buyers, investors, and developers the data-driven clarity they need in a volatile market. Moving forward, the goal is to expand this framework to cover the rest of Malaysia, perhaps integrating real-time economic indicators to make the predictions even more sensitive to sudden market shifts.

6. ACKNOWLEDGEMENT/FUNDING

Sincere appreciation is extended to Zaidah Ibrahim for her guidance and expertise in Machine Learning, and to the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, for the support in facilitating this work.

7. CONFLICT OF INTEREST STATEMENT

The authors declare that this research was conducted without any personal, commercial, or financial conflicts and report no conflicts of interest with the funders.

8. AUTHOR'S CONTRIBUTIONS

Muhammad Syazani Mohamad Jaffar: Conducted the research, performed data analysis and model development, and prepared the initial draft of the manuscript. Noor Latiffah Adam: Contributed to improving and refining the manuscript. Sharifalillah Nordin: Assisted the lead author in the preparation, review, and revision of the manuscript.

REFERENCES

Annamalai, R., Deena, S., Venkatakrishnan, R., & Harini, K. (2023). Automating machine learning model development: An OperationalML approach with PyCARET and Streamlit. 2023 Innovations in Power and Advanced Computing Technologies (i-PACT), 1-6. https://doi.org/10.1109/i-pact58649.2023.10434389

Huynh, T., Mazumdar, H., Gohel, H., Emerson, H., & Kaplan, D. (2023). *Evaluating the Predictive Power of Multiple Regression Models for Groundwater Contamination using PyCaret-23489* (No. INIS-US--24-WM-23489). WM Symposia, Inc., PO Box 27646, 85285-7646 Tempe, AZ (United States). https://inis.iaea.org/records/fkhy4-v3n70

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). Machine learning for property price prediction and price valuation: a systematic literature review. *Planning Malaysia*, *19*. https://doi.org/10.21837/pm.v19i17.1018

Li, Z. (2024). A Comparative Study of Regression Models for Housing Price Prediction. Transactions on Computer Science and Intelligent Systems Research. 5. 810-816. https://doi.org/10.62051/qjs7y352

https://doi.org/10.24191/mij.v7i1.11568

Mantaw, C. S., & Prahadeeswaran, M. (2024). Leveraging machine learning for accurate groundnut price forecasting in tamil nadu: an XGBOOST approach. https://doi.org/10.21203/rs.3.rs-4293571/v1

Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced machine learning techniques for predictive modeling of property prices. Information, 15(6), 295. https://doi.org/10.3390/info15060295

Mohd, T., Masrom, S., & Johari, N. (2019). Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 542–546. https://doi.org/10.35940/ijrte.b1084.0982s1119

National Property Information Centre (NAPIC). (2025). *Property market report Q1 2025: Residential overhang and transaction trends*. Ministry of Finance Malaysia. https://napic.jpph.gov.my/

Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In *2018 International conference on machine learning and data engineering (iCMLDE)* (pp. 35-42). IEEE. https://doi.org/10.1109/icmlde.2018.00017

Quispe, J. O. Q., Quispe, A. C. F., Calvo, N. C. L., & Toledo, O. C. (2024). Analysis and selection of multiple machine learning methodologies in pycaret for monthly electricity consumption demand forecasting. *Materials Proceedings*, *18*(1), 5. https://doi.org/10.3390/materproc2024018005

Tandon, R. (2024). The machine learning based regression models analysis for house price prediction. *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours*, *1*(3), 156-170.

Valuation and Property Services Department. (2024). Laporan Pasaran Harta 2023 [Property market report 2023]. Ministry of Finance Malaysia.

Wu, J. (2024). Multiple Machine Learning Models in House Price Prediction: Performance Evaluation and Comparison. Highlights in Business, Economics and Management. 40. 364-371. https://doi.org/10.54097/bd459218

Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: a survey of literature. *International Journal of Modern Education and Computer Science*, *12*(6), 46-54. https://doi.org/10.5815/ijmecs.2020.06.04

https://doi.org/10.24191/mij.v7i1.11568