REAL ESTATE PRICING USING MACHINE LEARNING AND TIME SERIES

N. Rikhotso and S.N. Ngidi

ABSTRACT

Machine learning (ML) models have improved real estate pricing by capturing complex nonlinear relationships, yet they typically overlook temporal dynamics such as market cycles and persistence. This study investigates whether incorporating these dynamics through hybrid modelling improves residential property price prediction. Using transaction data for one-family dwellings in Brooklyn, New York City, from 2006 to 2025, four ML models — Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Machine (SVM) and Gradient Boosting (GB) — are combined with an Autoregressive Integrated Moving Average (ARIMA) model applied to their residuals. The results demonstrate that hybrid models consistently outperform their standalone machine learning counterparts by producing lower prediction errors and substantially higher explanatory power. Residual modelling successfully captures persistent temporal patterns that are not explained by machine learning algorithms alone, resulting in significant improvements in forecasting accuracy across all evaluated models. While the SVM provides the strongest standalone performance, hybridisation improves the predictive accuracy of every model, with the largest relative gains observed for MLR and the highest overall predictive performance achieved by the SVM–ARIMA hybrid. Across all models, hybridisation reduces prediction error by roughly 35% while increasing the average coefficient of determination (R^2) from approximately 0.37 to 0.63. Diebold–Mariano tests further confirm that these improvements are statistically significant. These findings demonstrate that residuals contain valuable temporal information and that hybrid machine learning–time series frameworks provide a more accurate and robust approach to residential real estate valuation in dynamic housing markets.

KEYWORDS

Temporal Dependence; Residual Modelling; ARIMA; Market Dynamics; Predictive Performance.

CONTACT DETAILS

Ndzhaka Rikhotso, Student at the University of the Witwatersrand, ndzhakarikhotso11@gmail.com

Simama N. Ngidi, Student at the University of the Witwatersrand, simamangidi@gmail.com

1. INTRODUCTION

Real estate forms a large portion of household and institutional wealth in many countries, with a valuation of approximately $111 trillion across 34 developed countries, equivalent to almost 196% of their combined Gross Domestic Product (GDP) (Organisation for Economic Co-operation and Development, 2025). Beyond its contribution to national wealth, the real estate sector plays a

fundamental role in economic development through property taxation, construction activities, mortgage lending, insurance, brokerage services and investment markets (Organisation for Economic Co-operation and Development, 2025). Consequently, accurate real estate valuation is not merely an academic objective but a critical requirement for financial stability, efficient capital allocation, effective risk management, and informed decision-making by investors, financial institutions, insurers, and policymakers. Historically, real estate valuation has relied on hedonic pricing models, which estimate property prices as a function of observable characteristics such as size, location, age, floor level, and amenities (Rosen, 1974). These models provide an intuitive and theoretically grounded framework for understanding how individual property attributes contribute to market value. However, hedonic models are constrained by their linear functional form and the assumption that the marginal contribution of each characteristic remains constant over time. These assumptions become increasingly restrictive in dynamic housing markets such as New York City (NYC), where economic conditions, interest rates, government policies and buyer preferences continuously reshape the value placed on property characteristics despite the characteristics themselves remaining unchanged. Recent advances in data availability and computational techniques have shifted attention from traditional hedonic approaches towards machine learning (ML) models, motivated by their ability to capture complex nonlinear relationships and interactions that conventional statistical models often overlook (Baldominos, Blanco, Moreno, Iturrarte, Bernárdez & Afonso, 2018). Park and Bae (2015) demonstrate that ML models produce more accurate property valuations than traditional hedonic models using housing data from the United States. Subsequent studies by Baldominos et al. (2018) and Choy and Ho (2023) reached similar conclusions using data from Spain and Hong Kong respectively, providing consistent evidence that ML techniques outperform conventional hedonic models across different real estate markets. Despite their superior predictive performance (Park & Bae, 2015; Baldominos et al., 2018; Choy & Ho, 2023), ML models possess an important limitation: they generally treat each property transaction as an independent observation. Consequently, they fail to explicitly account for temporal dynamics such as serial correlation, market momentum, persistence, cyclical behaviour, policy interventions and structural changes that have long been recognised as fundamental characteristics of housing markets, particularly in rapidly evolving markets such as NYC (Glaeser & Gyourko, 2006). As a result, although ML models effectively capture cross-sectional variation between properties, they may overlook systematic temporal information embedded within historical transaction data. To address this limitation, recent research has begun integrating machine learning with time-series (TS) models through hybrid modelling frameworks that combine the strengths of both approaches. Among the earliest studies, Zhao (2024) demonstrated that incorporating temporal dynamics through a hybrid Random Forest–ARIMA framework improves property price prediction using Taiwanese housing data. These findings suggest that temporal information contained within machine learning residuals can be exploited to enhance forecasting accuracy. However, Zhao's (2024) study evaluates only a single hybrid specification, leaving unanswered whether hybridisation consistently improves different machine learning algorithms and whether certain hybrid combinations outperform others.

This study addresses these gaps by developing and comparing both standalone machine learning models and residual-based hybrid machine learning–time series models for residential property valuation. Specifically, the study is guided by three interrelated objectives. The first is to develop standalone ML and hybrid ML–ARIMA models for residential property price prediction. The second is to determine whether hybrid models consistently outperform standalone machine learning models in forecasting residential property prices. The third is to compare the predictive performance of different hybrid models to identify the framework that provides the highest forecasting accuracy. In addressing these objectives, the study makes both financial and actuarial contributions.

From a financial perspective, improved property valuation supports more informed investment decisions, mortgage lending, portfolio management and risk assessment while reducing the likelihood of valuation errors that may contribute to financial instability, as illustrated during the 2007–2008 global financial crisis (Packer & Riddiough, 2012). From an actuarial perspective, more accurate property price prediction provides insurers with improved estimates of replacement values and insured losses, thereby supporting more appropriate pricing of homeowners' and commercial property insurance products and reducing the risks associated with under-pricing or over-pricing insurance portfolios (Kim, Mahajan & Wang, 2024). The remainder of this paper is organised as follows. Section 2 reviews the relevant literature on hedonic pricing, machine learning, time-series forecasting and hybrid modelling approaches. Section 3 describes the data and methodology adopted in the study. Section 4 presents and discusses the empirical results, comparing the performance of standalone machine learning models with their hybrid counterparts. Finally, Section 5 concludes the paper by summarising the principal findings, discussing their implications, acknowledging the study's limitations and outlining directions for future research.

2. LITERATURE REVIEW

This section is organised into four parts. First, it reviews hedonic pricing models, focusing on their theoretical foundations, applications within real estate pricing, and their key strengths and limitations. Second, it reviews ML approaches, discussing the rationale for their increasing adoption, their performance relative to traditional hedonic models, and the challenges associated with their application. Third, it evaluates time series approaches, highlighting their primary advantage in capturing temporal dynamics as well as their key limitation when applied to real estate markets. Finally, it explores hybrid models by reviewing the existing literature, assessing the limited empirical evidence supporting their effectiveness and identifying the gaps that remain within current research.

Lancaster (1966) provides the theoretical foundation for the application of hedonic models to product pricing by arguing that consumers do not derive utility from products themselves, but from the characteristics of products. This reframes product pricing, moving from pricing a product as a single, indivisible good to pricing it as the sum of implicit prices placed by consumers on its individual characteristics. In the real estate context, this suggests that a property's price is determined by the sum of implicit prices placed by consumers on its observable characteristics, such as size, floor level,

location and the number of bedrooms. Rosen (1974) tests this theory in practice by developing a hedonic pricing model that regresses real estate prices on observable characteristics, showing that the theory holds empirically. Hedonic models offer several advantages that explain their decades-long dominance in real estate pricing. First, they are theoretically grounded (Lancaster, 1966; Rosen, 1974), making them credible. Second, they are computationally straightforward (Rosen, 1974), making it easy for individuals without specialist knowledge to use them to price real estate. Third, they produce interpretable results (Rosen, 1974), making it easy for individuals without specialist knowledge to understand results without any assistance. Fourth, they rely on actual market transaction data (Rosen, 1974), making them market consistent. Despite these advantages, however, they still have disadvantages that render them problematic in dynamic markets like NYC. First, they assume a linear relationship between the price of real estate and the implicit prices of its individual characteristics (Ekeland, Heckman & Nesheim, 2004), implying that they do not capture the complex non-linear relationships that characterise dynamic real estate markets like NYC (Soltani & Lee, 2024). Second, they assume that the marginal contribution of each characteristic to the property's price remains constant over time (Pakes, 2003), implying that they do not capture changes in economic conditions and shifting consumer preferences. Third, they treat each transaction as an independent observation (Choy & Ho, 2023), implying that they do not capture temporal dynamics such as serial correlation, momentum and mean-reversion.

The limitations of hedonic models motivated a shift from using hedonic models to using ML models to price real estate (Baldominos et al., 2018; Choy & Ho, 2023). This is because ML models are able to uncover hidden patterns in large real estate data sets, yielding lower prediction errors and better generalisation compared to hedonic models, as Park and Bae (2015), Baldominos et al. (2018) and Choy and Ho (2023) demonstrate. Park and Bae (2015) demonstrate this by developing a housing price prediction model based on ML algorithms (C4.5, RIPPER, Naïve Bayesian, and AdaBoost) using housing data from the US and finding that ML approaches achieved higher predictive accuracy compared to traditional hedonic models. Baldominos et al. (2018) demonstrate this by evaluating four ML approaches (Ensemble Methods, K-Nearest Neighbours (KNNs), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP)) using real estate data from Spain and finding that ensemble methods achieved the lowest Mean Absolute Error (MAE), implying that they outperform hedonic models in real estate pricing. Choy and Ho (2023) demonstrate this by comparing Extra Trees (ETs), KNNs and RF using housing transaction data from Hong Kong and finding that all three ML approaches achieved higher explanatory power (R^2) and lower prediction errors compared to hedonic models, implying that ML approaches outperform hedonic models in real estate pricing. While ML approaches outperform hedonic models, they still have limitations, like hedonic models. First, they may lack interpretability (McCluskey, McCord, Davis, Haran & McIlhatton, 2013), implying that it may be difficult, if not impossible, to understand exactly how they produced the value (McCluskey et al., 2013). Second, they treat observations as independent, overlooking temporal dynamics that are well documented in real estate markets (Glaeser & Gyourko, 2006; Choy & Ho, 2023).

While ML models treat transactions as independent, overlooking temporal dynamics, TS models can capture temporal dynamics (Crawford & Fratantoni, 2003; Miles, 2008). Crawford and Fratantoni (2003) demonstrate this capability by comparing the forecasting performance of three univariate TS models (Autoregressive Integrated Moving Average (ARIMA), Generalised Autoregressive Conditional Heteroskedasticity (GARCH) and the regime-switching class of models) with the intention of confirming their argument that regime-switching models are theoretically compelling for the real estate market. However, the results of their study reveal that while regime-switching models successfully capture temporal dynamics in-sample, simple ARIMA models generally perform better out-of-sample. Following Crawford and Fratantoni (2003), Miles (2008) demonstrates this capability using a slightly different approach. Miles (2008) argues that univariate models are insufficient for capturing boom–bust cycles in real estate markets and proposes a TS model (Generalised AutoRegression (GAR)) that accounts for non-linear dynamics in the real estate market. Miles (2008) then compares the out-of-sample forecast performance of GAR against that of ARIMA and GARCH models, finding that GAR performs best in volatile markets where boom–bust cycles are most pronounced, while it offers no improvement in relatively stable markets. Despite these demonstrations that TS models provide an advantage in capturing temporal dynamics, TS models still have one key limitation: they operate at the aggregate market level, neglecting the heterogeneity that exists in the real estate market (Crawford & Fratantoni, 2003).

The limitations of ML and TS models motivate the need for hybrid models that combine the strengths of the two while mitigating their weaknesses. Zhao (2024) is one of the first researchers to examine real estate pricing using hybrid models. In his study, Zhao (2024) demonstrates that hybrid models improve the accuracy of real estate price predictions. Using real estate data from Taiwan, he does so by comparing the forecasting performance of four ML models (MLR, RF, GB and SVM) to that of a hybrid model that combines RF and ARIMA models. His study shows that the hybrid model achieves a higher R^2 than all four ML models, suggesting that the hybrid performs best among the models compared. While these results provide encouraging evidence that hybrid models improve real estate pricing accuracy, the study has one key limitation that limits its generalisability and practical application: it tests only one hybrid model against four ML models, raising questions about whether other possible hybrid models would also outperform machine learning models and which hybrid model performs best among the different possible combinations.

3. DATA AND METHODOLOGY

3.1. DATA

In this study we focus on the residential real estate market of Brooklyn, one of the five boroughs of New York City. Among the five boroughs, Brooklyn is the most populous, with a population of approximately 2.9 million, a position owed largely to its geographic location (see Figure 1). Historically this location supported industrial development, which in turn attracted residents in search of economic opportunity; today the borough spans one of the widest ranges of residential submarkets

in the United States, from brownstone districts commanding multi-million-dollar prices to outer neighbourhoods of modest detached homes. This breadth, combined with two decades of pronounced market cycles, makes Brooklyn a demanding and informative setting in which to test pricing models.

Figure 1. Location of Brooklyn relative to New York City, New York State, and the United States.

The data are the property sales records published by the New York City Department of Finance, which records and maintains sales across all five boroughs for property tax assessment purposes. Our study covers twenty years of Brooklyn transactions, January 2006 to December 2025: 484,079 raw records across 21 fields. Each record describes the location of the property (address, neighbourhood, zip code, borough, block, and lot), its physical characteristics (building class category, building class at time of sale, tax class at time of sale, year built, residential unit counts, commercial unit counts, land square feet, and gross square feet), and the transaction itself (sale price and sale date). A field-by-field glossary is provided by the Department of Finance, and two definitions matter for what follows. First, gross square feet measures the total floor area of all structures from the exterior walls, so genuine vacant land carries a value of zero. Second, a sale price of zero denotes a transfer of ownership without cash consideration (a gift, inheritance, or intra-family deed transfer) rather than a market transaction.

Preliminary inspection revealed the data-quality issues typical of administrative records. Formatting was inconsistent in the categorical fields (the building class category alone contained 141 distinct strings arising from spacing variants, a spelling error, and duplicated naming conventions); the file contained records duplicated either exactly or on all fields except price; a large block of transactions carried prices of zero or a few dollars; properties classified as vacant land carried positive building areas and unit counts; unit counts, square footage, zip codes and construction years were missing or

zero for many records; and a small number of records carried construction years in the distant past. These issues motivated the systematic cleaning process summarised below.

Cleaning proceeded in three stages. The first stage standardised the categorical and text fields: neighbourhood names were trimmed, upper-cased and whitespace-collapsed; the building class category was reduced from 141 raw strings to 47 clean categories; and addresses were standardised with word-boundary pattern matching, with ordinal suffixes stripped so that duplicate detection is reliable.

The second stage removed unusable fields and records. Five columns were dropped: the borough identifier (constant), the easement field (entirely blank), the apartment number (71% missing), and the two "at present" classification fields, which describe the property today rather than at the moment of sale and would therefore inject future information into historical records (temporal leakage); the complete "at time of sale" classifications are used instead. Three invalid neighbourhood labels (125 records) and 83 records with a blank building class category were removed, as were 4,951 exactly duplicated rows and a further 139 rows duplicated on the key of block, lot, standardised address, sale date and sale price.

The third and most consequential stage concerned non-market transactions and internal contradictions. All sales below $10,000 were removed (182,993 records), consistent with standard practice in the New York housing literature: these are administrative records rather than market prices, and a model trained on them would learn that identical houses transact at both $0 and $900,000. Invalid zip codes were imputed with the modal zip of the record's neighbourhood. Properties whose sale-time building class codes (V0–V3) identify them as vacant land, corroborated by the category field, had contradictory building attributes set to zero. Missing unit counts were imputed with the modal value of the building class category and totals recomputed from their components. Missing or zero square footage was imputed with building-class medians, the median being robust to outliers, with the exception that condominium and co-operative unit sales legitimately record zero land area (the deed conveys a unit, not land) and were flagged rather than imputed. Construction years of zero, and prior to 1800, were imputed with class medians. Every imputed cell carries a binary indicator so that models can discount repaired values.

The overall cleaned dataset contains 295,788 arm's-length transactions across 38 columns with no missing values and full monthly coverage of 2006–2025; engineered fields include calendar decompositions, building age at sale, log-transformed square footages and sale prices, density ratios, and unit-sale indicators.

Table 1 gives an overview of the row accounting:

Cleaning step	Rows removed	Rows remaining
Raw Department of Finance file	—	484,079
Invalid neighbourhood labels	125	483,954

Blank building class category	83	483,871
Exact duplicate rows	4,951	478,920
Key duplicates (block, lot, address, date, price)	139	478,781
Non-arm's-length sales (price < $10,000)	182,993	295,788

Table 1. Row accounting for the data-cleaning pipeline.

From the cleaned data we study one-family dwellings (building class category 01), for two reasons. The first is homogeneity: a pricing function estimated across warehouses, condominium units and vacant land conflates fundamentally different price mechanisms. The second is economic relevance: one-family homes are the dominant owner-occupied asset and the natural object of valuation for households, lenders and insurers. The restriction leaves 39,660 transactions. Two modelling filters follow: sales below $100,000 (residual non-market or partial-interest transfers in this segment) and sales above the subset's 99.5th percentile ($6,400,000) are removed, amounting to 681 records, as are 12 records with physically implausible recorded areas. The final modelling sample contains 38,967 transactions.

The target variable is the natural logarithm of the sale price, reflecting the strong right skew of housing prices; model predictions are exponentiated back to dollars before evaluation, so all reported metrics are in US dollars. Thirty-three predictors are used, organised in four groups. The first group holds the structural characteristics of the property: log gross and log land square footage, the square of log gross square footage, building age at sale and its square, floor-area ratio, residential and commercial unit counts, a commercial-presence indicator, pre-war (built before 1945) and new-build (five years old or newer) indicators, and three imputation flags. The second group captures location at three spatial scales through smoothed target encoding (Micci-Barreca, 2001), in which each category is replaced by a shrunken average of the log price within it: zip code, neighbourhood, and tax block, the last of which places each sale within its immediate micro-location of roughly a city block. A fourth encoding is applied to the interaction of neighbourhood and detailed building class, allowing the value of a given house type to differ by area. To prevent information leakage, all four encodings are learned on the training window only and, within the training window, are computed out-of-fold. The third group holds relative-size features: the z-score of a property's log gross and log land area within its neighbourhood, computed from training-window statistics, which expresses whether a house is large or small for its area. The fourth group interacts the neighbourhood encoding with log gross and log land square footage, allowing the implicit price of floor space to vary with neighbourhood value, and one-hot encodes the detailed building class at time of sale across its twelve levels (A0–A9, S0, S1).

3.2. METHODOLOGY

Our methodology proceeds in ten steps. In the first step, the data are split chronologically into a training set (2006–2017) and a test set (2018–2025). A chronological split is essential rather than optional in this framework: a random split would scatter future transactions through the training data,

contaminating both the hedonic estimates and the residual series with information a real forecaster could never possess.

In the second step, four machine-learning models are trained on the training set, with hyperparameters and feature subsets selected on an internal validation split: a random 15% of the training window is held out, candidate configurations are fitted on the remainder and scored on the held-out portion, and the winning configuration is refitted on the full training window. Because the feature set is strictly cross-sectional, a random (rather than chronological) internal split is appropriate at this stage; the chronological separation of Step 1 is never crossed. MLR is estimated by ordinary least squares on standardised features and serves as the hedonic benchmark. Variable selection was performed for the linear model with the lasso (Tibshirani, 1996): the L1 penalty, with its strength chosen by five-fold cross-validation, retained 29 of the 33 predictors, discarding only features made redundant by the interaction terms (log land square footage, the raw neighbourhood encoding, squared log gross square footage, and one building-class indicator). The pruned model, refitted by ordinary least squares on the selected variables, validated marginally worse than the full specification, so the full model was retained; the selection exercise thus served to confirm the relevance of the feature set rather than to reduce it. The RF uses 400 trees, with the validation grid selecting a minimum leaf size of five and half of the features considered at each split. The SVM is implemented as a Nyström kernel-approximated support-vector regression, with 1,500 radial-basis components followed by a linear support-vector regressor ($C = 1$, $\varepsilon = 0.05$); the validation grid over the number of components, the kernel width and the regularisation constant confirmed this configuration. This implementation choice is important: exact kernel SVR scales between quadratically and cubically with sample size and is intractable on 23,000 observations, while the usual workaround (subsampling) discards data, and support-vector regression is highly sensitive to specification, since without feature standardisation the radial-basis kernel is dominated by the largest-magnitude inputs, and on a raw dollar target the ε-insensitive loss treats a $50,000 error identically on a $300,000 and a $3,000,000 property. Both standardisation and the log target are therefore applied. Finally, GB uses histogram-based boosting, with the validation grid selecting a learning rate of 0.05, 63-leaf trees and an L2 penalty of 1.0, and the number of iterations (roughly 260) set by early stopping on the validation split and rescaled for the full training window. All randomness is fixed (seed 42) for reproducibility.

In the third step, predictive performance on the test set is measured by mean absolute error, root mean squared error, and the coefficient of determination:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| P_i - P_i^{\mathrm{pred}} \right| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(P_i - P_i^{\mathrm{pred}} \right)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(P_i - P_i^{\text{pred}}\right)^2}{\sum_{i=1}^{n} (P_i - \bar{P})^2} \tag{3}$$

where P_i is the actual sale price, P_i^{pred} the predicted price, and $\bar{P}$ the mean of actual prices in the test set.

In the fourth step, residuals are computed on the training window. Two refinements are applied relative to the naive definition. First, residuals are taken from out-of-sample predictions within the training window (out-of-bag predictions for the RF and five-fold out-of-fold predictions for the remaining models) rather than from models evaluated on their own training data. The distinction is critical for flexible learners: a RF reproduces its training data almost perfectly, so in-sample residuals are near zero and carry none of the systematic bias the model exhibits on unseen data, which is the quantity the time-series stage must forecast. Second, residuals are defined in logarithmic space,

$$\varepsilon_i = \ln(P_i) - \ln\left(P_i^{\text{pred}}\right) \tag{4}$$

because prediction errors in an appreciating market grow proportionally with price levels, so a multiplicative correction is the natural functional form.

In the fifth step, following Zhao (2024), the residuals are averaged within each calendar year to generate a yearly residual time series:

$$\bar{\varepsilon}_t = \frac{1}{n_t} \sum_{i \in \text{year}\, t} \varepsilon_i \tag{5}$$

where n_t is the number of transactions in year t, giving twelve training observations (2006–2017).

In the sixth step, an ARIMA(p, d, q) model is fitted to the yearly residual series by conditional-sum-of-squares estimation. Orders are selected by genuine forecast validation rather than in-sample information criteria alone: the candidate model is fitted to all but the final three years of the residual series, the held-out years are forecast, and the order minimising held-out RMSE (with AIC as a tie-breaker) is refitted to the full training series. A trend guard restricts the search to d = 1 whenever the residual series exhibits a statistically significant linear trend ($|t| > 2.5$), preventing long-horizon forecasts from a spuriously stationary specification.

In the seventh step, the fitted ARIMA is combined with each machine-learning model to form a hybrid, and in the eighth step hybrid predictions are computed on the test set:

$$P_i^{\text{hybrid}} = P_i^{\text{pred}} \cdot \exp(\hat{\varepsilon}_t) \tag{6}$$

where $\hat{\varepsilon}_t$ is the forecast residual for the year containing sale i. The forecast is fully out-of-sample throughout: the ARIMA is estimated once on training-window residuals and projected statically across all eight test years, with no re-estimation and no test-period information of any kind.

In the ninth step, hybrid performance is measured with the metrics of Equations (1)–(3), and, because a hybrid could outperform its standalone counterpart by chance, the improvement is tested formally with the Diebold–Mariano statistic under the Harvey–Leybourne–Newbold small-sample correction, computed on monthly-aggregated squared-error differentials over the ninety-six test months with a Newey–West long-run variance (twelve lags).

In the tenth step, the metrics of hybrid and standalone models are compared and conclusions drawn.

4. RESULTS AND DISCUSSION

4.1. MARKET CONTEXT AND RESIDUAL BEHAVIOUR

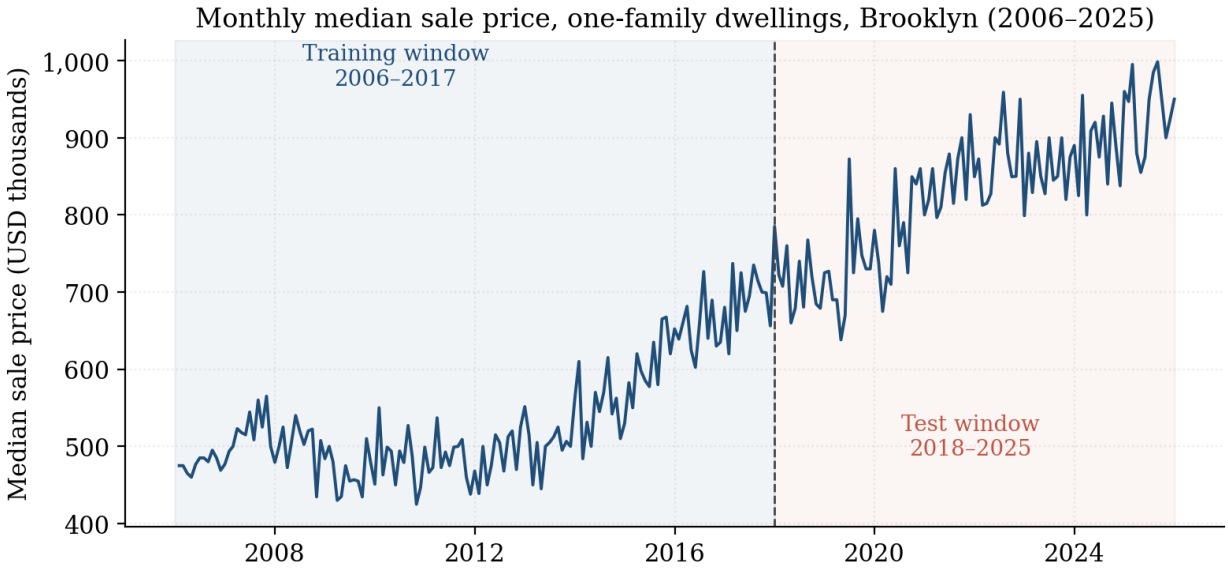

Figure 2. Monthly median sale price of one-family dwellings in Brooklyn, 2006–2025, with the chronological train/test split. Prices in the test window sit roughly 70% above the training-window average, appreciation that a time-blind hedonic model cannot see.

Figure 2 frames the forecasting problem. Median one-family prices roughly doubled over the sample, with the 2008–2009 downturn, the long post-2013 expansion, the brief pandemic disruption and the subsequent surge all visible; the test window sits far above training-period levels. Residual analysis plays a critical role in determining whether additional predictive information remains unexplained by a model. If residuals are purely random and centred on zero, there is little justification for further modelling; if they display trends, persistence or cyclical behaviour, the primary model has failed to capture information embedded in the data-generating process. Figure 3 presents the yearly average out-of-fold training residuals for the four models, the series the ARIMA stage must learn.

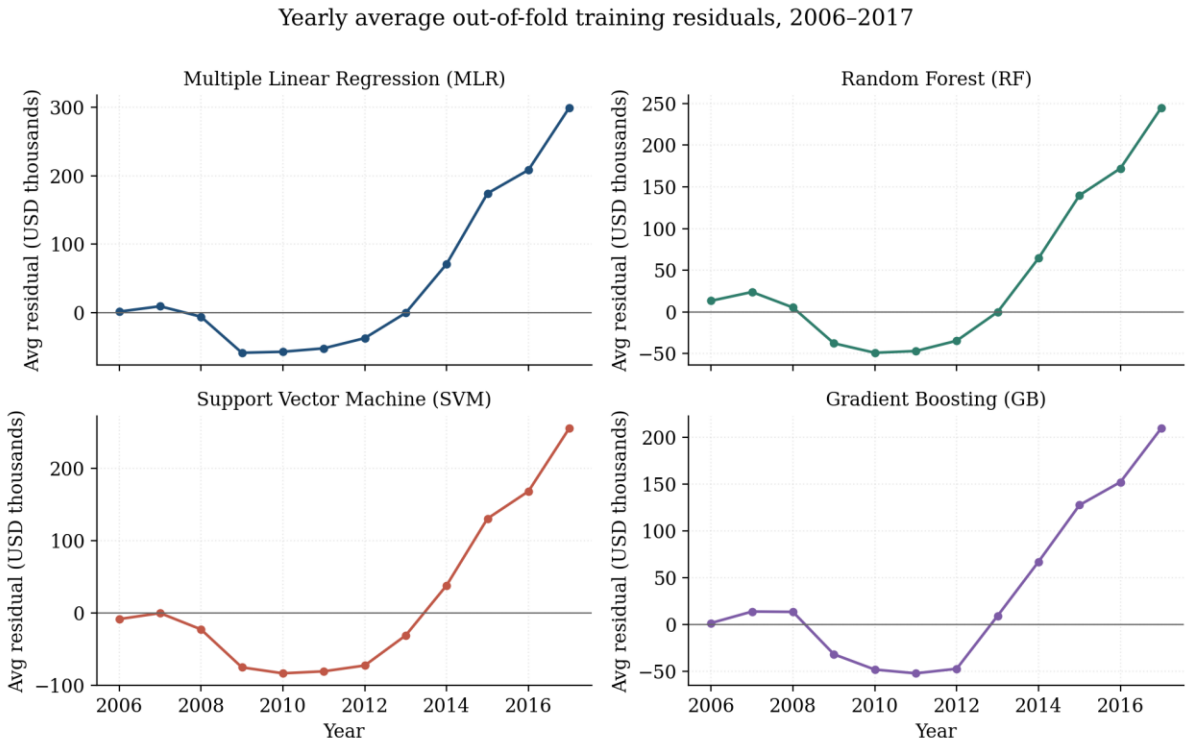

Figure 3. Yearly average out-of-fold training residuals (USD), 2006–2017, by model. The common trough marks the Global Financial Crisis; the subsequent drift is the appreciation signal exploited by the ARIMA stage.

All four series tell one story with different amplitudes. Residuals hover near zero through the mid-2000s, dip negative through the crisis years (the models, informed partly by later data under the out-of-fold scheme, over-predict during the trough), and then drift steadily and increasingly positive from 2013 onward as realised prices pull away from what a purely cross-sectional model can justify. The linear model shows the steepest drift, reflecting its rigid functional form; the tree ensembles and the SVM are more moderate but unambiguous. None of these series resembles white noise: the residuals are systematic, persistent and forecastable, which is the empirical premise of the hybrid framework.

4.2. STANDALONE MODEL PERFORMANCE

Model	MAE	RMSE	R^2
MLR	$468,809	$724,140	0.273
RF	$440,966	$671,959	0.374
SVM	$421,651	$649,255	0.416
GB	$442,180	$654,623	0.406

Table 2. Standalone machine-learning performance on the 2018–2025 test window (USD).

Among the standalone models the SVM performs best ($R^2 = 0.416$), followed by GB (0.406) and the RF (0.374), with the linear model weakest ($R^2 = 0.273$). Feature attributions from the RF indicate what carries the cross-sectional fit: the neighbourhood-by-building-class encoding is the single most influential predictor, followed by the block-level encoding and the interactions of neighbourhood value

with property size. In other words, the fine spatial structure of the market, down to the individual block and to how a given house type prices within its area, does most of the work, and it is this structure, together with the kernel's ability to exploit the smooth interaction surface, that lifts the SVM to the top of the standalone ranking. All four models remain limited in absolute terms, and by design: they were denied every temporal signal while the test window sits roughly seventy per cent above training-period price levels. The purpose of Table 2 is therefore not to identify the best hedonic model but to establish the baseline that the time-series stage must repair, and to quantify how much of observed price variation is temporal rather than cross-sectional in origin.

4.3. THE TIME SERIES STAGE

For three of the four models (MLR, SVM and GB), the validation-based selection procedure chose ARIMA(0, 1, 0) with drift for the yearly log-residual series, that is, a random walk with drift; for the RF, whose residual series is closer to stationary fluctuation around a positive trend over the validation window, an ARIMA(1, 0, 1) specification was selected instead. The dominance of the drift specification is itself informative. With twelve annual observations and an eight-year static forecast horizon, the reliably estimable signal is the drift term: the models' proportional under-prediction grows at a stable estimated rate, and richer autoregressive structure rarely survives held-out validation. Figure 4 provides the central validation of the framework: for each model, the ARIMA forecast, estimated purely from 2006–2017 residuals and projected eight years ahead with no re-estimation, is plotted against the residuals the standalone model actually produced across 2018–2025. The realised series track the forecasts closely and remain within the 95% forecast intervals. The systematic under-prediction the standalone models develop over the test window is, to a first approximation, the path the residual models predicted.

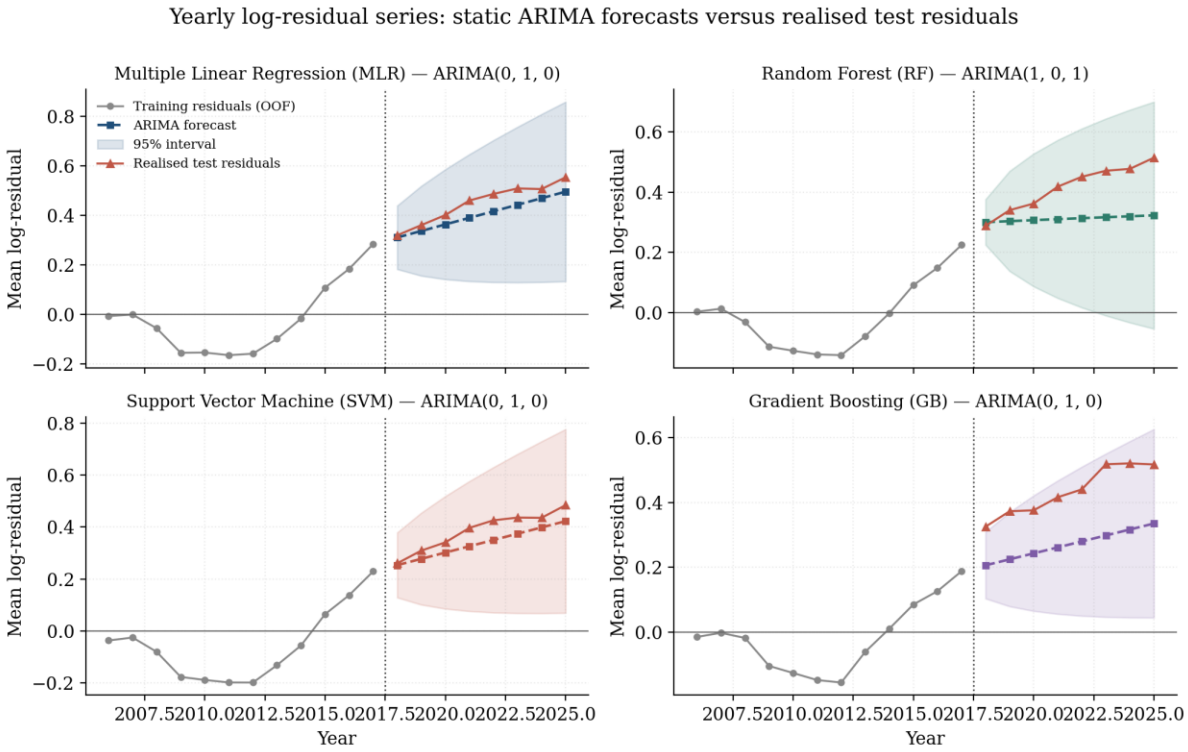

Figure 4. Yearly log-residual series by model: training residuals (out-of-fold), static ARIMA forecasts with 95% intervals, and realised test-window residuals.

4.4. HYBRID MODEL PERFORMANCE

Model / variant	MAE	RMSE	R^2	DM p-value
MLR — standalone	\$468,809	\$724,140	0.273	—
MLR — hybrid, Eq. (6), ARIMA (0, 1, 0)	\$281,898	\$531,583	0.608	< 0.001
RF — standalone	\$440,966	\$671,959	0.374	—
RF — hybrid, Eq. (6), ARIMA (1, 0, 1)	\$287,478	\$508,036	0.642	< 0.001
SVM — standalone	\$421,651	\$649,255	0.416	—
SVM — hybrid, Eq. (6), ARIMA (0, 1, 0)	\$275,154	\$504,472	0.647	< 0.001
GB — standalone	\$442,180	\$654,623	0.406	—
GB — hybrid, Eq. (6), ARIMA (0, 1, 0)	\$307,662	\$515,358	0.632	< 0.001

Table 3. Standalone versus hybrid performance (test window, USD). DM p-values test each hybrid against its standalone counterpart on monthly squared-error differentials (HLN-corrected).

Three findings stand out. First, every hybrid outperforms its standalone counterpart on every metric (the central hypothesis of this paper), and the Diebold–Mariano tests reject equal predictive accuracy at any conventional level for all four models ($p < 0.001$): the improvement is systematic, not fortuitous. Second, the gains are large: averaged across the four models, MAE falls by roughly 35% and R^2 rises from 0.37 to 0.63. Third, the ordering of the hybrids mirrors the standalone ordering: the SVM hybrid

is the best configuration in the study (R^2 = 0.647, MAE $275,154), while the largest absolute improvement accrues to the weakest baseline (MLR, ΔR^2 = 0.335): the effectiveness of residual correction scales with the degree of systematic structure in a model's errors, so the more systematic the forecasting bias, the greater the benefit derived from modelling it. Figures 5 and 6 summarise the comparison.

Performance Comparison: Standalone vs Hybrid Models (Sale Price in USD)
Hybrid = ML + ARIMA on yearly log-residuals · DM = Diebold–Mariano test vs standalone

Algorithm	MAE Standalone	MAE Hybrid	MAE Improve	RMSE Standalone	RMSE Hybrid	RMSE Improve	R^2 Standalone	R^2 Hybrid	R^2 Improve	DM p-value
MLR vs MLR_Hybrid	$468,809	$281,898	↓ 39.9%	$724,140	$531,583	↓ 26.6%	0.273	0.608	↑ 0.335	< 0.001
RF vs RF_Hybrid	$440,966	$287,478	↓ 34.8%	$671,959	$508,036	↓ 24.4%	0.374	0.642	↑ 0.268	< 0.001
SVM vs SVM_Hybrid	$421,651	$275,154	↓ 34.7%	$649,255	$504,472	↓ 22.3%	0.416	0.647	↑ 0.231	< 0.001
GB vs GB_Hybrid	$442,180	$307,662	↓ 30.4%	$654,623	$515,358	↓ 21.3%	0.406	0.632	↑ 0.226	< 0.001

Figure 5. Performance comparison of standalone and hybrid models, with relative improvements and Diebold–Mariano significance.

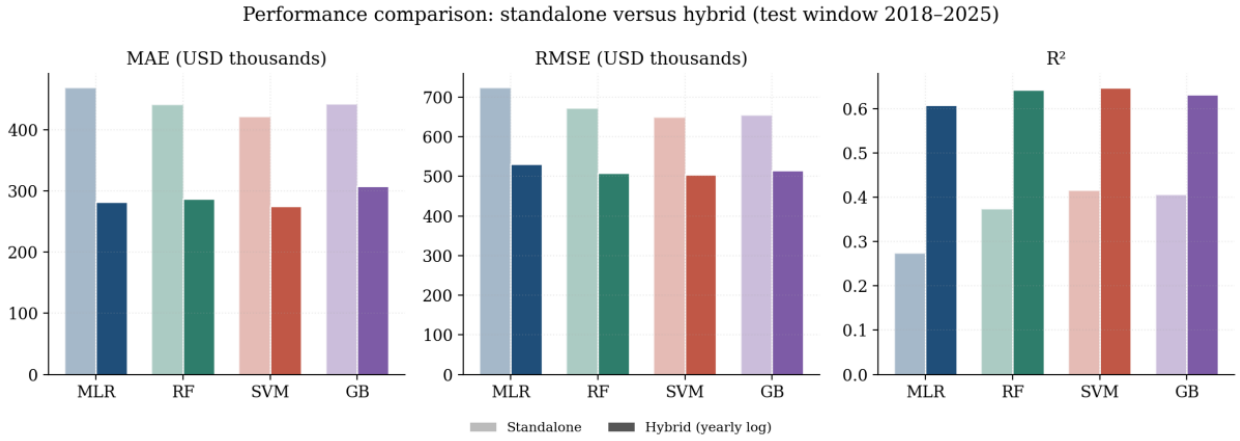

Figure 6. Performance comparison: standalone versus hybrid across the three evaluation metrics.

4.5. PREDICTION DIAGNOSTICS

Figures 7–9 examine the predictions themselves. In Figure 7 the standalone predictions sit systematically below the 45-degree line (the visual signature of the temporal level gap), while the hybrid correction re-centres every model's cloud without materially changing its shape: the ARIMA stage repairs the level, and the hedonic stage retains responsibility for cross-sectional ordering. Figure 8 reports Regression Error Characteristic (REC) curves, the regression analogue of ROC analysis, plotting the share of predictions within a given relative error tolerance; the hybrid curve dominates its standalone counterpart at every tolerance for every model, and the area-under-curve summaries roughly double. At the practically relevant 20% tolerance, the hybrid models price roughly three-fifths

of transactions within tolerance, against roughly one-fifth for the standalones. Figure 9 traces mean absolute error across the test years: standalone error grows steadily as the market appreciates away from training-period levels, while the hybrid holds error broadly flat across the full eight-year horizon, which is direct evidence that the drift correction, estimated once in 2017, remains serviceable through 2025.

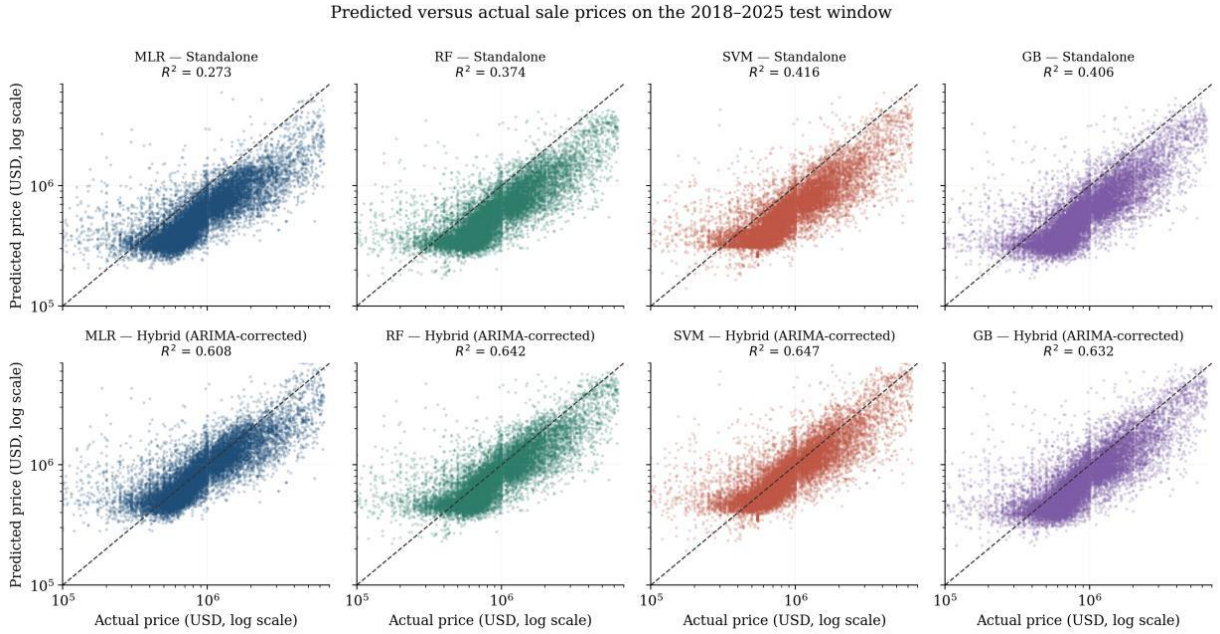

Figure 7. Predicted versus actual sale prices on the test window (log scales). Standalone predictions (top row) sit systematically below the 45° line; the hybrid correction (bottom row) re-centres them.

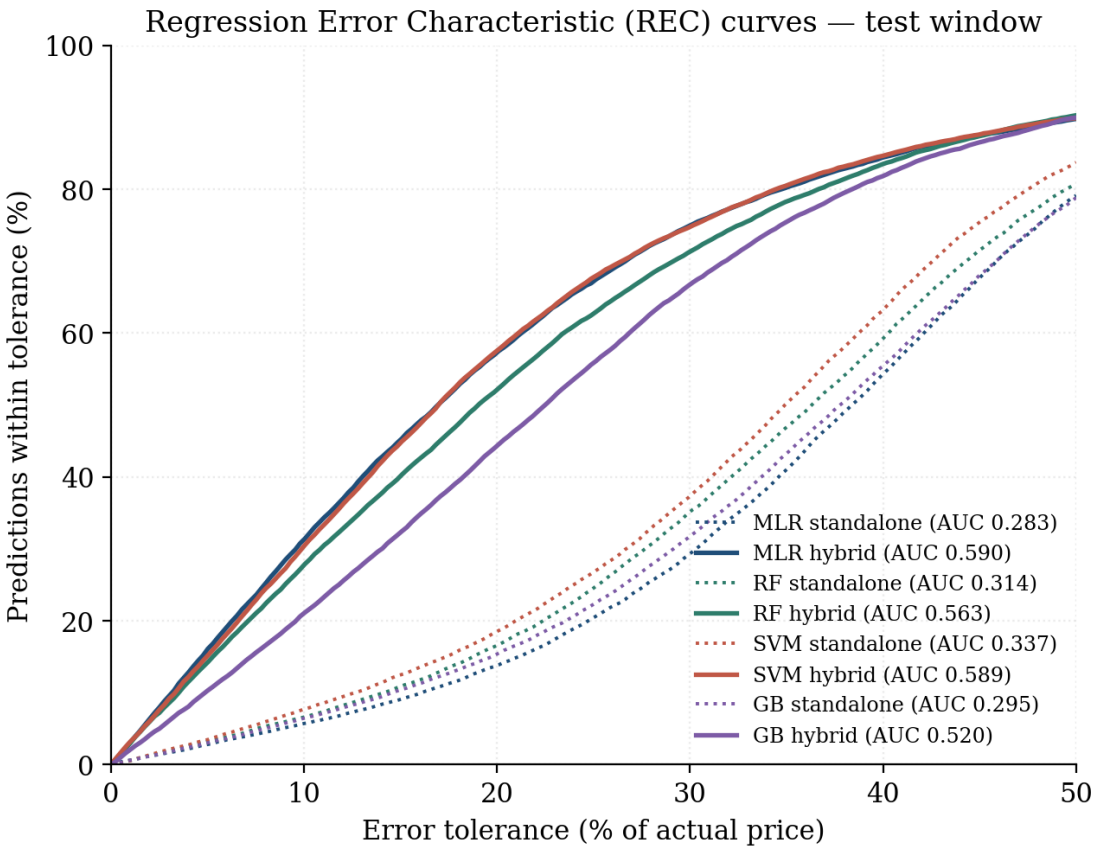

Figure 8. Regression Error Characteristic curves. For every model the hybrid (solid) dominates the standalone (dotted) at every tolerance; area-under-curve values in the legend.

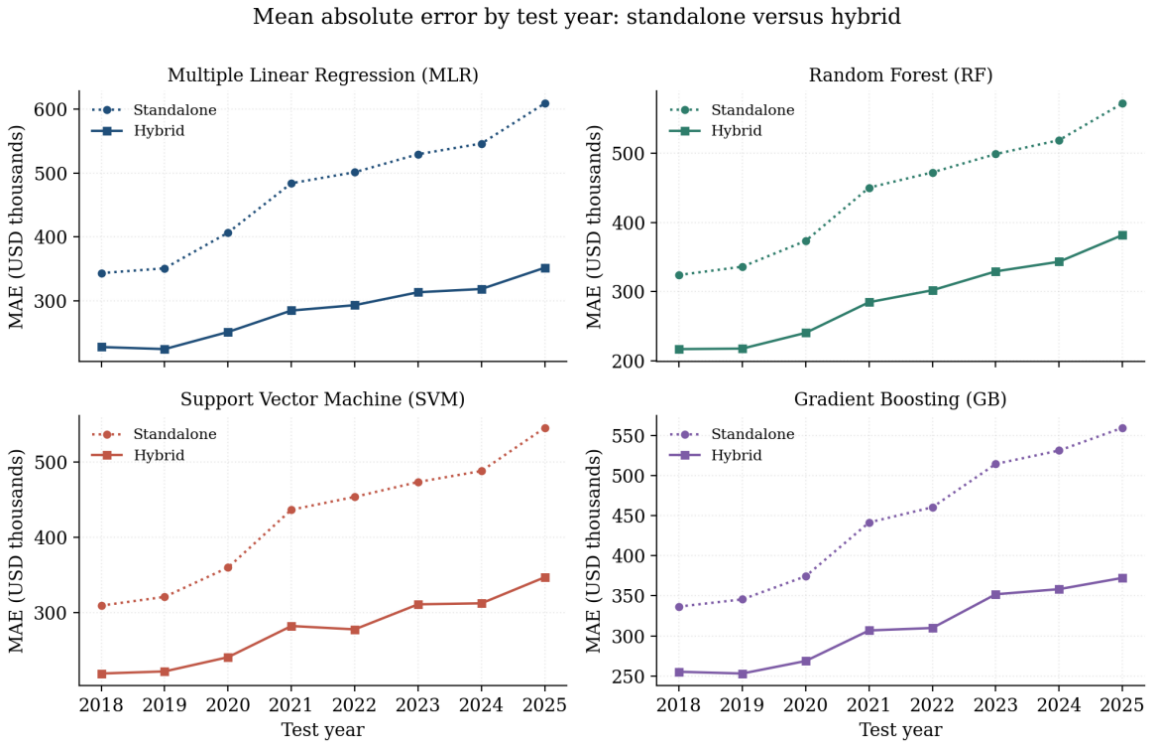

Figure 9. Mean absolute error by test year: standalone versus hybrid.

5. CONCLUSION

This study set out with three interrelated aims: to develop hybrid and machine-learning models for real estate pricing; to determine whether hybrid models consistently outperform standalone machine-learning models; and to identify which hybrid configuration performs best. Using twenty years of one-family dwelling transactions in Brooklyn (2006–2025) under a strict temporal separation (cross-sectional machine learning trained on 2006–2017 with no time variables, an ARIMA stage fitted to yearly averages of out-of-fold training residuals, and a single static projection across an eight-year test horizon), the answers are clear. Hybrids outperform their standalone counterparts for all four algorithms examined, on every metric, with Diebold–Mariano tests rejecting equal accuracy at $p < 0.001$ in each case; averaged across the four models, mean absolute error falls by roughly 35% and R^2 rises from 0.37 to 0.63. The best configuration is the SVM hybrid ($R^2 = 0.647$, MAE \$275,154), and the data selected a random-walk-with-drift residual model for three of the four algorithms, implying that the dominant temporal signal is a stable proportional drift.

The study makes several contributions. Methodologically, it demonstrates that residuals generated by machine-learning models should not be treated as random noise: they contain systematic, forecastable temporal structure (in this market, a stable proportional drift arising from appreciation that a cross-sectional model cannot see), and modelling that structure delivers large, statistically verified gains. Beyond confirming the promise of hybridisation identified by Zhao (2024), the paper contributes refinements of independent value in both stages. In the machine-learning stage, location information at three spatial scales (zip code, neighbourhood and tax block), a neighbourhood-by-building-class interaction encoding, neighbourhood-relative size features and interactions of neighbourhood value with property size substantially strengthen the cross-sectional fit, with hyperparameters chosen by internal validation and variable selection by the lasso confirming the relevance of the feature set. In the time-series stage, residual series are built from out-of-fold rather than in-sample predictions, without which flexible learners supply the ARIMA stage with noise; the correction is multiplicative in log space, matched to the proportional nature of pricing errors; and ARIMA orders are chosen by forecast validation with a trend guard, which reduces long-horizon instability. Empirically, the study examines a large and economically significant market over a horizon spanning the Global Financial Crisis, the post-crisis recovery, the COVID-19 pandemic and the subsequent monetary tightening cycle, and shows that the hybrid advantage is consistent across four algorithms of very different character (a linear model, two tree ensembles and a kernel method). Practically, the framework offers valuation professionals, mortgage lenders, investors and policymakers a forecasting architecture whose components are separately interpretable: a hedonic surface for the property and a drift term for the market. From an actuarial perspective, more accurate property values serve as a key input to pricing homeowners' and commercial property insurance, reducing the risk of mispricing that leads to financial losses or diminished competitiveness for insurers (Kim, Mahajan & Wang, 2024); more broadly,

accurate collateral valuation bears directly on the financial-stability concerns highlighted by the 2007 crisis (Packer & Riddiough, 2012).

The broader implication of this study is that relying on a single modelling paradigm is insufficient for complex economic systems such as housing markets, where cross-sectional structure and temporal dynamics coexist and neither reduces to the other. This study demonstrates that a deliberate division of labour, machine learning for the pricing surface and time-series analysis for the market's movement over time, outperforms either alone, consistently, significantly, and across every algorithm we considered. As data availability and analytical methods continue to advance, frameworks that combine machine learning, econometrics, time-series analysis and spatial modelling are likely to define the frontier of property price forecasting; the evidence assembled here adds to the growing case that such hybrid approaches are the productive direction for both the accuracy and the practical utility of real estate valuation.

REFERENCES

Baldominos, A., Blanco, I., Moreno, A.J., Iturrarte, R., Bernárdez, Ó. & Afonso, C., 2018. Identifying real estate opportunities using machine learning. Applied Sciences, 8(11), 2321. https://doi.org/10.3390/app8112321

Choy, L.H.T. & Ho, W.K.O., 2023. A machine learning-based comparative analysis of property price prediction models. Land, 12(4), 740. https://doi.org/10.3390/land12040740

Crawford, G.W. & Fratantoni, M.C., 2003. Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. Real Estate Economics, 31(2), 223–243. https://doi.org/10.1111/1540-6229.00064

Ekeland, I., Heckman, J.J. & Nesheim, L., 2004. Identification and estimation of hedonic models. Journal of Political Economy, 112(S1), S60–S109. https://doi.org/10.1086/379947

Glaeser, E.L. & Gyourko, J., 2006. Housing dynamics. NBER Working Paper No. 12787. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w12787

Kim, M., Mahajan, P. & Wang, Z., 2024. The rise in insurance costs for commercial properties: Causes, effects on rents, and the role of owners. SSRN Working Paper No. 5175451. https://ssrn.com/abstract=5175451

Lancaster, K.J., 1966. A new approach to consumer theory. Journal of Political Economy, 74(2), 132–157. https://doi.org/10.1086/259131

McCluskey, W.J., McCord, M., Davis, P.T., Haran, M. & McIlhatton, D., 2013. Prediction accuracy in mass appraisal: A comparison of modern approaches. Journal of Property Research, 30(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

Micci-Barreca, D., 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter, 3(1), 27–32. https://doi.org/10.1145/507533.507538

Miles, W., 2008. Boom–bust cycles and the forecasting performance of linear and non-linear models of house prices. The Journal of Real Estate Finance and Economics, 36(3), 249–264. https://doi.org/10.1007/s11146-007-9067-1

Organisation for Economic Co-operation and Development (OECD), 2025. Future-proofing real estate investment: Place-based risks. OECD Urban Studies. Paris: OECD Publishing. https://doi.org/10.1787/2dd12063-en

Packer, F. & Riddiough, T., 2012. Securitisation and the commercial property cycle. In: RBA Annual Conference Volume 2012: Property Markets and Financial Stability. Sydney: Reserve Bank of Australia. Available at: https://www.rba.gov.au/publications/confs/2012/packer-riddiough.html

Pakes, A., 2003. A reconsideration of hedonic price indices with an application to PC's. American Economic Review, 93(5), 1578–1596. https://doi.org/10.1257/000282803322655455

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. https://doi.org/10.1016/j.eswa.2014.11.040

Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55. https://doi.org/10.1086/260169

Soltani, A. & Lee, C.L., 2024. The non-linear dynamics of South Australian regional housing markets: A machine learning approach. Applied Geography, 166, Article 103248. https://doi.org/10.1016/j.apgeog.2024.103248

Zhao, S., 2024. Applying machine learning and time series to predict real estate valuations. In: Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2024). Amsterdam: Atlantis Press, pp. 479–486. https://doi.org/10.2991/978-94-6463-490-7_52