Multi-Model Architecture Evaluation for Amenity Based Price Prediction in the Kingston Region

Dolev Klein Harari
Queen's University
23gjd@queensu.ca

Prerana Sil
Queen's University
prerana.sil@queensu.ca

Barberry Yu
Queen's University
24xn30@queensu.ca

Ivan Bardziyan
Queen's University
22tm57@queensu.ca

Noah Premasinghe
Queen's University
23flw2@queensu.ca

***Abstract*—Municipal urban planners often face difficulties in determining the optimal location for developing city infrastructure due to the enormous number of contributing factors in evaluating economic ramifications. In collaboration with the City of Kingston we determined property valuations to be strong indicators of the economical activity of an area. We introduce a dynamic multi model architecture evaluation to assist our client, The City of Kingston, in developing a spatial econometric modeling system for urban planning purposes. Our approach evaluates four machine learning models trained on residential property transaction data using location-based attributes as predictor variables. Spatial clustering and logarithm functions were incorporated to capture non-linear relationships commonly observed in housing prices. Overall, our success metrics demonstrate that location-based factors play a significant role in determining residential property values, particularly in university-centered cities like Kingston. This project provides municipal staff with a practical tool to help narrow-down and refine planning scenarios, helping improve efficiency and strengthen evidence-based policy development in urban planning.**

I. Introduction

A. Motivation

The valuation of homes is important for the decision-making of municipalities, which seeks to understand housing trends to guide planning decisions and infrastructure strategies. This project is a collaborative effort with the City of Kingston's Strategy, Innovation & Partnerships department to build spatial econometric models that can forecast how relocating a development can impact property values. The aim of this project is to act as a tool for the City of Kingston for integration in the city's planning workflow.

Traditional housing valuation methods often heavily rely on structural property characteristics, such as unit size or age, to predict house prices. While these structural attributes are important determinants in predicting house values, our study aims to further examine the influence of location-based factors on the value of the property. This is particularly relevant in Kingston, a university-centered city with more than 30,000 students in a population of around 150,000 residents. As a result, factors such as proximity to campus, healthcare services, social amenities, and green spaces play an increasingly important role in predicting housing values.

To evaluate the influence of locational features, we apply predictive modeling techniques to estimate current housing prices. This evaluation focuses on present-day price estimation rather than long-term depreciation or property aging effects. Through this approach, we aim to better understand the influence of location-based external factors and neighbourhood wealth on residential property valuations in Kingston.

B. Related Works

Aydin et al. [6] developed predictive models using gradient linear regression (GLR), geographic weighted regression (GWR), and forest-based classification and regression (FBCR) to evaluate comparative performance. Their results indicate that GWR and FBCR outperform alternative approaches, primarily due to their ability to effectively account for spatial heterogeneity by capturing non-linear relationships within the data.

Copiello [7] examined spatiotemporal dependence in residential property values in Northeastern Italy. Instead of focusing on singular house values, the study aggregates housing values to investigate serial and spatial autoregressive models. A logarithmic transformation was applied on the dependent variable to counteract non-linearity in housing values. Findings indicate that serial lag was much more significant than spatial lag, with regression coefficients of spatial lags illustrating that its effect decreases as distance grows. To account for this limitation, our research focuses specifically on the City of Kingston (451.58 square kilometers), allowing us to minimize the distance-related effects and to examine spatial lag as the core mechanism for relocation valuation.

Sharm et al., [8] aims to develop a house price prediction model and compares ML techniques to ascertain the best ML model. Results show that XG boosts are the most suitable regression technique with a strong R-squared value (0.920) and the lowest root mean squared error (RMSE 0.112). Validation was further confirmed with a histogram representing residual plot for the XGBboost model where distribution seemed to follow a bell-shape curve. GridSearchCV hyperparameter tuning also proved to be beneficial for enhancing model performance.

Cellmar et al., [9] conducted a study in three Polish cities to investigate the significance of various POI categories on housing prices. Ten categories of POIs connected with the functioning of urban space were proposed: sustenance, education, transportation, healthcare, entertainment, public service and financial, facilities, shops and services, leisure and sport, tourism, and historical. The Ordinary Least Squares regression was the technique used to determine coefficients of linear regression equations. The model ultimately revealed that only

three POI categories may be necessary for the formation of housing prices: sustenance, leisure and tourism (p-value was lower than 0.05). Overall, there was a statistically significant link between the number of POIs nearby and the prices of residential properties in each city.

C. Problem Definition

1) Contextual Information: This study is completed with the intention of dynamically predicting economic ramifications of relocating infrastructure in the City of Kingston. In collaboration with the municipality, we agree that housing valuations are an effective method with which to evaluate economic impacts, and therefore models developed in this study would be used to evaluate the changes in price given that certain POIs are moved. Housing valuation data is sourced from the Municipal Property Assessment Corporation, an organization which completes housing inspections for municipal property tax. Due to this, the study aims not to determine the market price of properties - as those are highly reliant on the national housing market - but rather on the fundamental value of the property as seen by the government.

2) Assumptions:

- Distance to points of interest is the primary quantitative measurement to represent location
- Different locations of each point of interest (ie. various parks) carry equal weighting.
- The distance of one kilometer in the center of the city functions the same as one kilometer in the suburbs.
- Kingston is self-contained and not influenced by neighbouring municipalities (ie. Frontenac County, Greater Napanee).

3) Restrictions: The scope of this work is subject to several constraints that define its applicability and scope of use.

- *Geographic specificity:* All models were trained on data Kingston-Specific data, therefore, it is not expected that models will generalize to other municipalities without retraining on local data sources.
- *Temporal limitations:* 2021 Census data was used in training, and therefore might not generalize fully to the modern state unless retrained.
- *Property type scope:* The dataset that was used to train the models is exclusively based on residential properties and does not include commercial, industrial, mixed-use, or rural properties. Therefore, predictions on property types other than residential areas are not within the scope of this work.
- *POI data completeness:* The features for calculating distance-to-POI are derived from a fixed set of amenities retrieved from the City of Kingston open data portal. Any changes in the set of amenities after the date of collection will not be reflected in the features and could lead to stale information in inference results.
- *Neighborhood boundary coverage:* SQFT Features were derived from polygon boundaries, not actual metric size, thus might not reflect data fully.

II. Methodology

This study evaluates the performance of four machine learning models in determining the value of Kingston properties. The models are proven through previous literature to demonstrate high accuracy in the property evaluation domain, and thus are chosen to provide tangible benchmarks for our study. The models tested are Decision trees, Neural Networks, Random Forests, and XGBoost. In prior research, these models have shown promising performance due to their innate capability to work with non-linear relationships in real estate data.

A. Data Engineering

1) Dataset Construction: In constructing the completed training set, data were compiled from a series of databases open sourced by the City of Kingston (Figure 1). These datasets were merged on the *Address* key to form the final dataset containing features seen in Table I.

Points of Interest
POI Name
POI Type
Metric Coordinates

Property Tax Assessments
Address
Price
Metric Coordinates
Unit Type

2021 Neighborhood Census
Neighborhood Name
Mean Income
GeoJSON Boundaries
Tax Bracket Distribution

Buildings
Address
Storeys
Shape Area
Metric Coordinates

Fig. 1. Database schema for the four datasets used in this study.

These datasets are used to engineer our complete feature set Table I.

2) Distance Metrics: Given that locational properties of POIs and residential properties are described in the Universal Transverse Mercator (UTM) coordinate system, Euclidean distance was used to derive distance metrics.

Euclidean Distance:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

Although Haversine distance is generally more accurate for geographic coordinates, the limited spatial span of the dataset makes Euclidean distance sufficiently precise.

This produces an $N \times M$ distance matrix between properties and POIs. To reduce dimensionality, distances were averaged across similar POI classes, yielding a 23-dimensional feature space.

3) Additional Features: SQFT and Neighborhood Wealth were derived from The "Buildings" and "2021 Census" databases respectably. Square footage was estimated as:

$$\text{SQFT} = \text{STOREYS} \times \text{SHAPEAREA} \tag{2}$$

Where SHAPEAREA is the area of a polygon representing the unit in digital mapping software.

TABLE I
DATABASE SCHEMA FOR THE FOUR DATASETS USED IN THIS STUDY.

Feature
POI Distances
NEIGHBOURHOOD_WEALTH
sqft

POI Distances SubFeatures
Airport
Bus Transit
Childcare
Civic
Community Services
Corrections
Cultural
Emergency
Ferry
Healthcare
Hotel
Industrial
Infrastructure
Marina Water Access
Parking
Parks
Postsecondary
Religious
School
Shopping Centre
Sports Recreation
Trails
Train Transit
Unmapped

4) Target Variable Processing: Exploratory analysis (scatterplots, correlation matrices, and price distribution maps) indicated that the target variable Price (P) exhibited high skew and outliers.

Outliers were filtered using an IQR filter:

$$Q_1 - 1.5\,IQR \leq P \leq Q_3 + 1.5\,IQR \tag{3}$$

Price was transformed using either a log transform:

$$P' = \log(P) \tag{4}$$

or a Box–Cox transformation:

$$P^{(\lambda)} = \begin{cases} \dfrac{P^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(P), & \lambda = 0 \end{cases} \tag{5}$$

The Box–Cox transformation optimizes λ to stabilize variance and reduce skew.

5) Locational Price Smoothing: Due to a lack of structural features (bathrooms, bedrooms, yard size, etc.), variance among similarly located properties is reduced using a K-Nearest Neighbors smoothing approach.

For each property i, its locational price LPr_i is defined as:

$$LPr_i = \frac{1}{K} \sum_{j \in \mathcal{N}_K(i)} P_j \tag{6}$$

where $\mathcal{N}_K(i)$ denotes the set of K nearest neighbors of property i. The target property's own price was excluded to avoid data leakage.

6) Amenity Density Feature: A final feature was constructed: *n_amenities_1km* representing the count of amenities within a 1 km radius, intended to help differentiate rural and urban areas.

B. Decision Tree Model

A Decision Tree was selected due to its versatility in capturing non-linear relationships and was developed using the Scikit-learn Python library.

The dataset was partitioned into training and test sets using an 67/33 random split.

Following Train-test split, the Decision Tree regressor was implemented with a maximum tree depth of 14. Due to Decision Trees being prone to overfitting given unbounded depth, 14 was found, through hyperparameter tuning, to allow the model to capture complex relationships between housing attributes and prices while limiting excessive growth that may lead to overfitting.

C. Random Forest Model

The Random Forest Regressor model was trained using Scikit-Learn's `RandomForestRegressor` [10] to predict Kingston residential property prices. When implementing the train_test_split, we used a test_size of 33%. Similarly to the Decision Tree model, Random Forests are prone to overfitting. A Grid Search was run to identify optimal hyper parameters to minimize squared error, and returns optimal parameters:

- n_estimators=500
- random_state=42
- min_samples_leaf=3
- n_jobs=-1
- bootstrap=True
- criterion="squared_error"
- max_depth= 100

Random Forest Performance Plots are seen in Figure 2.

D. Neural Network

A fully connected feedforward neural network was developed using PyTorch to perform regression over Kingston residential property prices. The architecture, training procedure, and feature engineering pipeline are described below.

1) Feature Engineering: The final feature set comprised 35 input variables grouped into two categories: base features, and neural network specific engineered features.

The 26 base features are seen in Table I, and X and Y coordinates from Figure 1.

All distance features were log-transformed prior to modelling to linearise the diminishing-returns relationship between proximity and price.

The 7 neural network-specific engineered features were:

- `local_house_price_mean`: Derived using equation 6 where K=20.
- `avg_essential_dist`: The mean distance to healthcare, school, and emergency POIs.
- `avg_transit_dist`: The mean distance to bus, train, and ferry POIs.
- `avg_recreation_dist`: The mean distance to sports, parks, and trail POIs.
- `wealth_per_avg_dist`: The ratio of neighbourhood wealth to mean POI distance, a composite desirability signal.

- `log_wealth`: log-transform of neighbourhood wealth, included to model diminishing marginal effects.
- `corrections_minus_parks`: signed difference between distance to corrections facilities and distance to parks, encoding the contrast between undesirable and desirable proximity effects.

All features were standardised to zero mean and unit variance using a standard scaler fitted exclusively on the training partition, with the fitted scaler serialised to disk and applied consistently at inference time.

The dataset was partitioned into training (70%), validation (15%), and test (15%) sets using stratified random splits with a fixed random seed (42) to ensure reproducibility across all models in this study.

2) Model Architecture: The neural network is a multi-layer fully connected (dense) regression network. The architecture consists of four hidden layers, with widths $[256, 128, 64, 32]$, followed by a single linear output neuron with no activation function. Each hidden layer applies the following sequence of operations:

$$h^{(l)} = \text{Dropout}\left(\text{ReLU}\left(\text{BatchNorm}\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)\right)\right) \tag{7}$$

Batch normalization is included after each linear transform to stabilize the training dynamics across the wide range of input scales, and Dropout with probability $p = 0.2$ was applied after each activation to act as a regularizer and reduce overfitting. The final layer maps the 32-dimensional representation to a single scalar output $\hat{y}$ representing the predicted log-price. The total number of trainable parameters is approximately 44,000.

3) Training Procedure: The network was trained by minimising Mean Squared Error (MSE) over log-transformed prices:

$$L = \frac{1}{N}\sum_{i=1}^{N} N\left((\hat{y}_i - \log\left(1 + p_i\right)\right)^2 \tag{8}$$

The Adam optimiser was used with an initial learning rate of $\eta = 1 \times 10^{-3}$ and L2 weight decay of $\lambda = 1 \times 10^{-5}$. A learning rate scheduler halved the learning rate whenever the validation loss did not improve for 10 consecutive training epochs, allowing the optimiser to converge into finer minima in later stages of training. Training was capped at 200 epochs with early stopping applied if the validation loss showed no improvement for 15 consecutive epochs. The model checkpoint corresponding to the lowest observed validation loss was restored before the evaluation. Training was conducted with a batch size of 256.

4) Evaluation: Model performance is assessed on the held-out test set using three metrics reported in original dollar scale: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Permutation-based feature importance was computed post training by measuring the increase in test mean square error (MSE) (in log-price space) when each feature was independently shuffled across five repetitions, providing an interpretable ranking of feature contributions without modifying the trained model weights.

E. XGBoost Model

A gradient boosted regressor was developed using the XGBoost Python Package to predict residential property prices in Kingston. An XGBoost regressor model was chosen given the success demonstrated in previous literature predicting the valuation of residential properties. The dataset preproessing, architecture, training, and model evaluation are described below.

1) Data Preprocessing: The model was trained on the dataset outlined by the data construction section (II-A1).

2) Model Architecture and Hyperparameter Tuning: The XGBoost regressor was configured with 900 estimators, a maximum tree depth of 7, and a learning rate of 0.028. The slow learning rate was chosen to reduce the risk of over shooting key parameters during boosting iterations. To mitigate overfitting L1 and L2 regularization were incorporated ($\alpha = 0.3, \lambda = 2.2$). The stronger L2 weight was used to promote generalization. A minimum child weight of two was used to prevent overfitting of small leaves.

3) Evaluation: Model performance was assessed on both training and test sets using four metrics: MAE, RMSE, R^2, MAPE.

III. Results

TABLE II
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	MAPE	RMSE	MAE	R^2
Decision Tree	17.8%	76,022.84	48032.58	0.36
Random Forest	**10.83%**	**48,225.37**	**32136.34**	0.68
XGBoost	12.80%	71,496.81	44,297.03	**0.79**
Neural Network	16.50%	103,486.01	57,069.04	0.50

A. Model Analysis

Due to the shared tree based architecture of the decision tree, random forest, XGBoost models they each exhibit similar strengths and weaknesses across results. The XGBoost model achieves a R^2 of 0.7924 on the training set and 0.7239 on the test set respectively (Fig. 3), indicating a moderate degree of explanatory power. The training MAE of CAD$40,078.23 versus the test MAE of $44,297.03 corroborates this, once again reflecting reasonable but imperfect generalization to unseen properties. Similarly, both the random forest and decision tree model have a larger difference between coefficients of determination indicative of overfitting (Fig. 2) In the XGBoost model RMSE diverges more substantially between training ($61,975.89) and test ($71,496.81) which is consistent with the residual plot revelations of heteroskedasticity (Fig. 3). The model achieves a mean error of $850.73 and is extremely unbiased. The random forest model achieves the lowest RMSE of the four models tested at $48,225.37, and has less outlier predictions seen in the residual plot of Fig. 2.

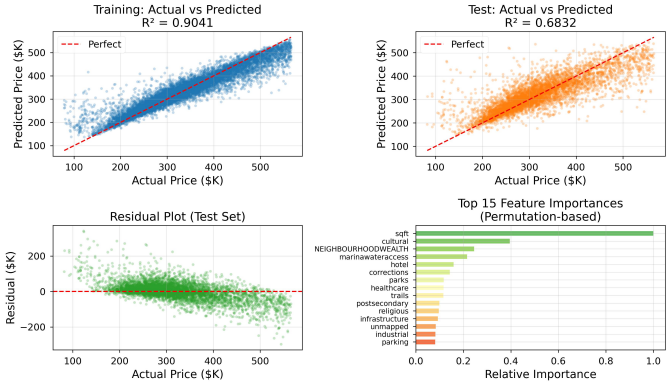

Fig. 2. Random Forest Plots

The residual plot from Figure 3 reveals that XGBoost the model struggles with predicting the value of more expensive houses, as shown by the increased spread from an actual value of $600,000 onward in the graph. This is a common feature of property valuation models.

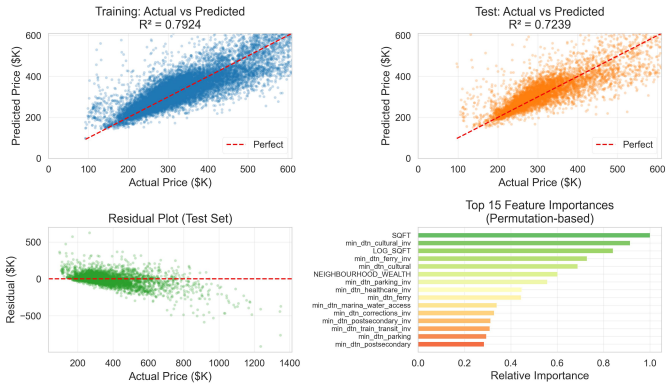

Fig. 3. XGBoost Evaluation Metrics

The residual plot shown in Figure 3 demonstrates extreme outliers, explaining contrast between MAE and RMSE.

Similarly, the Neural Network model seen in Figure 4 displays diminishing accuracy as Price increases, but in contrast, appears to perform significantly better on lower bound data than the tree based models.

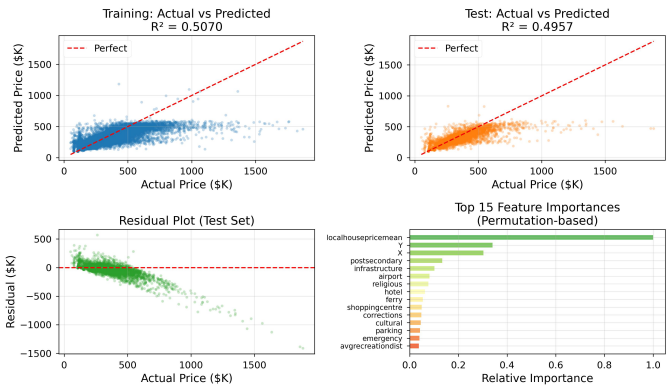

Fig. 4. Neural Network Evaluation Plots

B. Which Model was Best?

Of the four models evaluated, the Random Forest demonstrated the most robust overall performance. Although the XGBoost model achieved the highest R^2 (0.79), its MAE ($44,297) and RMSE ($71,496.81) are approximately 40% larger than the Random Forest's metrics. RMSE values for all models are relatively high; since RMSE penalizes the large errors quadratically, this discrepancy suggests that models are on average doing worse on edge case data points. This might indicate a fundamentally different distribution once a certain price threshold is reached, and might be investigated further with a combination of multiple models trained on different data quadrants. The Random Forest produced the lowest RMSE across all models and an R^2 of 0.68, only 0.11 below the XGBoost's 0.79. Furthermore, as an ensemble method averaging over many de-correlated trees, the Random Forest is structurally less prone to overfitting than a single Decision Tree, making it the more generalizable and reliable model for deployment on unseen data.

C. Impacts

These models have provide utility for urban planners looking to audit optimal locations for development. By using models as described in our study, initial location audits can be fully automated to provide a shortlist of economically beneficial locations to manually explore. Additionally, this research provides insight to Municipality data storage practices, and highlights where further data can be collected. Finally, these models can be implemented with the City of Kingston to test efficacy in a real world infrastructure development workflow.

IV. Conclusions

A. Future Steps

This study produced multiple robust models providing statistically significant predictions on the relation between POIs and property values. Future steps in the project would include:

- Discussing effective data collection and storage practices with the City of Kingston; ensuring normalization across datasets, monitoring missing data, developing an API for developer usage.
- Retraining models on combined Structural & Locational data. It is possible that relationships between distance metrics and price are dependent on subsidiary structural metrics, and therefore it would likely increase model performance to expand the feature space to include this feature sphere.
- Integrating into an easily interpretable system to be used by urban planners to efficiently create initial recommendation for new infrastructure projects.

REFERENCES

[1] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," ScienceDirect, https://www.sciencedirect.com/science/article/pii/S0957417414007325 (accessed Mar. 2, 2026).

[2] H. Lee, H. Han, C. Pettit, Q. Gao, and V. Shi, "Machine Learning Approach to residential valuation: A convolutional neural network model for geographic variation - the annals of regional science," SpringerLink, https://link.springer.com/article/10.1007/s00168-023-01212-7 (accessed Mar. 2, 2026).

[3] "Kingston and Area Real Estate Association," Kingston and Area Real Estate Association — CREA Statistics, https://creastats.crea.ca/board/king/ (accessed Mar. 2, 2026).

[4] Q. Gao, V. Shi, C. Pettit, and H. Han, Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia, https://www.sciencedirect.com/science/article/pii/S0264837722004367 (accessed Jan. 12, 2026).

[5] H. Lee, H. Han, C. Pettit, Q. Gao, and V. Shi, "Machine Learning Approach to residential valuation: A convolutional neural network model for geographic variation - the annals of regional science," SpringerLink, https://link.springer.com/article/10.1007/s00168-023-01212-7 (accessed Mar. 4, 2026).

[6] "Predict home prices with regression analysis and machine learning," Redirect to: /en/, https://learn.arcgis.com/en/projects/build-house-valuation-models-with-machine-learning/ (accessed Mar. 4, 2026).

[7] S. Copiello, Spatial dependence of housing values in Northeastern Italy, https://www.sciencedirect.com/science/article/pii/S0264275119302938 (accessed Oct. 3, 2025).

[8] H. Sharma, H. Harsora, and B. Ogunleye, "An optimal house price prediction algorithm: Xgboost," arXiv.org, https://arxiv.org/abs/2402.04082 (accessed Mar. 4, 2026).

[9] R. Cellmer, M. Be, and R. Trojanek, "Housing prices and points of interest in three Polish cities - journal of housing and the built environment," SpringerLink, https://link.springer.com/article/10.1007/s10901-024-10124-7 (accessed Mar. 4, 2026).

[10] "Randomforestregressor," scikit, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed Mar. 4, 2026).

[11] "XGBoost Documentation," https://xgboost.readthedocs.io/en/release_3.2.0 (accessed Mar. 3, 2026).