សាកលវិទ្យាល័យនរតុន
NORTON UNIVERSITY
Educating Tomorrow's Leader

Master of Science in Information Technology

Graduate School

PROPERTY VALUE PREDICTION USING MACHINE LEARNING:

A HISTOGRAM GRADIENT BOOSTING APPROACH FOR PROPERTY VALUATION IN CAMBODIA

A Data-Driven Machine Learning Approach to Reliable Real Estate Price Estimation

Subject: Machine Learning

Lecturer: Sek Socheat

Group 10 : 1. Oum Vuthy 2. Chan Sovychin 3. Chan Reaksmey 4. Chhorn Menghong

Graduate School, Norton University, Phnom Penh, Cambodia

Machine Learning Course — Final Assignment

August 2026

Abstract

Cambodia's real estate market has expanded rapidly in recent years, yet property pricing across the country remains largely informal. Valuations are commonly set through agent intuition, word-of-mouth, and social-media asking prices rather than through a consistent, evidence-based process. This informal approach produces inconsistent and sometimes inflated prices, widens the information gap between buyers and sellers, and slows mortgage and collateral assessments for banks and lenders. This paper presents a machine-learning-based Automated Valuation Model (AVM) for Cambodian residential property, developed as a final Machine Learning course project at Norton University. A dataset of 1,001 property listings, collected across Cambodian cities and districts, was cleaned and transformed through currency and numeric cleaning, text-based feature extraction, geographical parsing, categorical classification, and dimensionality reduction. After filtering price and land-size outliers and engineering a district-level price indicator, a final modelling dataset of 880 records and seven predictor features — bedrooms, bathrooms, land size, a unified building-area measure, district price level, city, and property type — was used to train a Histogram Gradient Boosting Regressor (HGBR). Numeric features were median-imputed and robust-scaled, categorical features were one-hot encoded, and the target price was log-transformed to reduce the influence of high-value luxury listings. The model was trained on 85% of the data and evaluated on a 15% held-out test set. On the test set, the model achieved an R^2 of 0.9373, a Mean Absolute Error of $34,617.38, a Mean Absolute Percentage Error of 6.01%, and produced predictions within ±20% of the true price for 96.97% of test listings. These results indicate that HGBR can capture the non-linear, location-dependent relationships that shape Cambodian property prices, with accuracy comparable to benchmarks reported for XGBoost and Random Forest models in other property markets. The resulting model offers a practical reference tool for buyers, sellers, agents, investors, and lenders, while its reliance on asking-price listings, a single-country dataset, and simplified outlier thresholds point toward clear directions for future refinement.

Keywords: *Machine Learning; Property Valuation; Real Estate; Histogram Gradient Boosting; Automated Valuation Model; Cambodia; Price Prediction.*

1. Introduction

1.1 Background

Cambodia's real estate sector has grown from a small, largely informal trade into a significant part of the national economy, driven by urban expansion, Borey (gated housing) developments, and a growing stock of condominiums and high-rise units in Phnom Penh and other provincial centres. Despite this growth, the way individual properties are priced has not kept pace with the scale of the market. Property valuation in Cambodia is still largely relationship-driven: prices are frequently set through agent intuition, informal negotiation, and word-of-mouth comparisons rather than through a documented, data-based process.

This informality creates a gap between how quickly the market is expanding and how reliably individual properties are priced. Sellers and agents often anchor asking prices to comparable listings shared on social media or property pages rather than to a statistical model of the underlying drivers of value, such as location, size, and legal title status. As Cambodia's financial sector becomes more digitised — for example through the Bakong payment system — the absence of an equally modern, data-driven approach to property valuation stands out as a gap that machine learning is well positioned to help close. Machine learning models can, in principle, learn the complex and non-linear relationships between a property's characteristics and its market price directly from historical listing data, offering a transparent and reproducible alternative to purely subjective appraisal.

1.2 Problem Statement

The project material that motivated this study identifies several structural weaknesses in how Cambodian property is currently valued:

- Manual and subjective valuation — prices are commonly set through relationship-based opinion rather than a systematic, evidence-based process.
- Information gap — limited transparency between buyers and sellers reduces market efficiency and makes it difficult for either side to verify a fair price.
- Emotional or inflated asking prices — “over-asking” behaviour driven by seller expectation rather than measurable market value.
- Banking and mortgage appraisal inefficiency — manual appraisal slows collateral assessment and loan-approval timelines for lenders.
- Missing and noisy property data — the source listings show meaningful gaps in physical-size fields, most notably in building area.
- Cambodian location-specific complexity — value is shaped by factors such as district, city tier, and title type that a simple average cannot capture.
- Extreme property-price outliers — a small number of very high-value “trophy” listings can distort simple averages and bias naive models.

It is worth noting that the project's own supporting materials are not fully consistent on the exact scale of the missing-data problem: one presentation slide reports approximately 17% missing land-size values, while a separate exploratory data analysis chart in the same project reports 0% missing land size and 38% missing building area. This paper reports both figures explicitly in Section 4.1 rather than silently resolving

the discrepancy, and treats the 38% missing building-area figure — which is consistent across the project material — as the better-supported estimate of data noise in the physical-size fields.

1.3 Research Gap

Traditional valuation approaches, whether based on agent judgement or simple comparable-sales averaging, struggle to represent the non-linear and interacting effects of property characteristics on price. A large land parcel does not add value at a constant rate once it exceeds a practically usable size; a given bedroom count is worth more in Phnom Penh than in a smaller provincial city; and legal title status can create a price premium that a simple location average cannot isolate. These are exactly the kinds of non-linear, interaction-heavy relationships that tree-based ensemble methods are designed to capture, which motivates the use of a machine-learning approach — specifically Histogram Gradient Boosting — for this project rather than a linear or purely rule-based valuation method.

1.4 Research Aim

The aim of this research is to develop a machine-learning-based Automated Valuation Model (AVM) capable of estimating Cambodian property prices from property characteristics and location-related information, using a Histogram Gradient Boosting Regressor trained on a cleaned, engineered dataset of Cambodian property listings.

1.5 Research Objectives

1. Build a Histogram Gradient Boosting model for Cambodian property-price prediction.
2. Collect and analyze Cambodian property data.
3. Clean, transform, and prepare the dataset for machine learning.
4. Produce accurate and reliable property-price estimates.

1.6 Research Questions

1. Which property features most influence property prices in Cambodia?
2. Can Histogram Gradient Boosting provide accurate property-price predictions?
3. How well does the model generalize across different locations and property types?

1.7 Significance of the Study

A reliable, data-driven valuation reference has practical value for a range of market participants. Home buyers and sellers gain a transparent reference price before entering negotiation. Real estate agents can support listings and client advice with data-backed estimates rather than opinion alone. Property investors can use the model to assess value and potential returns across different Cambodian cities and districts. Banks and lenders can use faster, more consistent estimates to inform mortgage and collateral assessments. Government bodies and researchers can draw on the underlying dataset and model outputs for housing-market analysis and policy planning.

2. Literature Review and Related Work

2.1 Traditional Property Valuation

Conventional property valuation relies primarily on expert or agent judgement, informed by comparable recent sales, general market familiarity, and negotiation experience. This comparable-sales approach can work reasonably well in stable, data-rich markets, but it depends heavily on the appraiser's access to accurate comparables and is difficult to standardise across appraisers, districts, or property types. In a market such as Cambodia's, where formal transaction records are limited and asking prices are often used as a proxy for value, traditional valuation is particularly exposed to inconsistency and subjective bias.

2.2 Machine Learning for Real Estate Prediction

Supervised machine learning reframes property valuation as a regression problem: given a set of property characteristics (features) and a set of properties with known prices (labels), a model learns a function that maps features to price. Unlike a simple comparable-sales heuristic, a supervised model can weigh dozens of interacting variables simultaneously and can be evaluated objectively against held-out data that the model has not seen during training. This makes machine learning attractive for markets where the relationship between property characteristics and price is complex, non-linear, and shaped by many interacting local factors.

2.3 Tree-Based Ensemble Models

Among supervised methods, tree-based ensembles have become a standard choice for structured, tabular data such as property listings. A single Decision Tree splits the data recursively on the feature values that best separate high- and low-price properties, but a standalone tree tends to overfit and generalise poorly. Random Forest reduces this overfitting by averaging many decision trees trained on bootstrapped samples of the data. Gradient Boosting methods, including XGBoost and Histogram Gradient Boosting, build trees sequentially, with each new tree correcting the errors of the ensemble built so far. Histogram Gradient Boosting Regressor (HGBR), the algorithm used in this project, further bins continuous features into histograms before splitting, which speeds up training on large tabular datasets and allows the algorithm to handle missing values natively without requiring a separate imputation step for every feature.

2.4 Automated Valuation Models

An Automated Valuation Model (AVM) is a system that produces a property-price estimate directly from structured data and a trained statistical or machine-learning model, without requiring a human appraiser to inspect each property individually. AVMs are already used by lenders and property platforms in many markets to provide fast, consistent, and reproducible price estimates. The Histogram Gradient Boosting model developed in this project functions as a prototype AVM for the Cambodian residential market: given a property's bedrooms, bathrooms, land size, building area, district, city, and property type, it returns a predicted price without manual appraisal.

2.5 Handling Missing and Noisy Property Data

Real-world property listings are rarely complete. Building-area and land-size fields are frequently missing or inconsistently reported, and a small number of extremely high-value listings can distort simple statistics such as the mean price. Modern AVMs must therefore be able to tolerate noisy, incomplete, high-dimensional data. This is one of the main reasons HGBR was selected for this project: it can natively handle missing values during training, and when combined with explicit outlier filtering and a log-transformed target, it is comparatively robust to the kind of data noise present in the Cambodian listings used here — including the missing building-area values described in Section 1.2.

2.6 Comparative Literature Table

Table 1 summarises the benchmark figures referenced in the project's supporting material alongside the result obtained in this study. Several of the cited studies are given in the source material only as an author name and year, without a full bibliographic reference (journal, volume, pages, or DOI); these entries are marked for verification in the reference list in Section 8 rather than presented with fabricated citation details.

Study / Model	Year	Algorithm	R^2	Error Metric	Key Finding
Al Marzooqi & Redouane [1]	2025	XGBoost	0.93	MAE: 16,347	Best overall balance of speed and precision across six supervised models compared.
Sevgen & Tanrivermiş [2]	2024	Random Forest	>0.90	MAPE: 8.8%	Highly stable performance for structured datasets across diverse urban contexts.
Zilli et al. [3]	2024	ANN vs. traditional HPM	N/A	MAPE: 16% vs. 38%	AI/ML model roughly halves prediction error compared with a traditional Hedonic Pricing Model.
Decision Tree (standalone) [4]	Not reported	Decision Tree	0.385	Not reported	More interpretable but markedly lower precision than ensemble methods.
This Project — HGBR	2026	Histogram Gradient Boosting	0.9373	MAPE: 6.01%	Comparable to elite benchmark accuracy despite Cambodia's noisy, incomplete listing data.

Table 1. Comparative summary of benchmark studies referenced in the project material and this project's HGBR result.

2.7 Research Gap and Contribution

The studies summarised in Table 1 confirm that ensemble tree-based methods consistently outperform both simple decision trees and traditional hedonic pricing approaches on structured real-estate data, generally achieving R^2 values above 0.90 in the markets where they have been studied. However, none of the referenced benchmarks were derived from Cambodian data, and the project material does not

identify a prior published AVM built specifically for the Cambodian residential market. This project's contribution is therefore twofold: first, it applies Histogram Gradient Boosting — a comparatively recent and computationally efficient ensemble method — to a real, noisy Cambodian property dataset rather than a clean benchmark dataset; and second, it documents a complete, reproducible preprocessing and modelling pipeline (Sections 3.2–3.10) tailored to the specific data-quality issues found in Cambodian listings, including missing building-area values and extreme price outliers.

3. Materials and Methods

3.1 Research Framework

The project followed a linear machine-learning workflow, moving from raw listing data to a trained, evaluated model. Table 2 summarises the stages of this workflow in the order they were implemented.

Stage	Description
1. Data Collection	Property listings gathered from public Cambodian real-estate listing sources.
2. Data Loading	Raw listings loaded into a working dataframe (1,001 initial records).
3. Data Cleaning	Currency/numeric cleaning, text-based size extraction, geographic parsing, categorical classification, and column pruning (Section 3.4).
4. Exploratory Data Analysis	Review of price distribution, missing-value patterns, and feature correlations (Section 4.2).
5. Feature Engineering	Construction of the unified Building_Area feature and the District_Price_Level location feature (Section 3.6).
6. Outlier Filtering	Removal of extreme price and land-size values (Section 3.5).
7. Preprocessing	Median imputation and robust scaling of numeric features; one-hot encoding of categorical features (Section 3.4).
8. Train/Test Split	85% training / 15% test split, random_state = 42.
9. HGBR Training	Histogram Gradient Boosting Regressor trained on a log-transformed target (Section 3.7–3.8).
10. Prediction	Price predictions generated for the held-out test set.
11. Evaluation	R^2, MAE, MAPE, and accuracy-within-±20% computed on the test set (Section 3.10).
12. Visualization	Actual-vs-predicted scatter plot generated to inspect prediction quality (Figure 4).

Table 2. Machine-learning workflow implemented in the project, from data collection through evaluation and visualization.

3.2 Dataset

The modelling dataset consists of Cambodian residential property listings compiled from public property-listing sources, in line with the project's data-collection plan, which identifies public listing platforms, web-scraped listings, and — where available — real estate agency and open government data as candidate sources. The listings cover residential property types (land, house, villa, and condominium) across Phnom Penh and a number of Cambodian provinces, with each record describing the property's physical characteristics, its location (village, district, and city), and its listed price. According to the implemented data-loading code, the working dataset begins with 1,001 initial records, and all 1,001 of these contained a valid Price_Cleaned value at the loading stage, meaning the dropna step applied to the price column removed no rows at that point in the pipeline. This differs from a separate exploratory-analysis figure in the project material, which reports approximately 1.6% missing price values; this paper reports the code-based figure as authoritative for the implemented pipeline, per the project's own instruction to prioritise

the final implemented code over descriptive presentation slides where the two differ, and flags the discrepancy here for transparency.

3.3 Data Features

After outlier filtering and feature engineering (Sections 3.5–3.6), the final model was trained on seven predictor features and one target variable, summarised in Table 3.

Feature	Type	Description	Role
Bedrooms	Numerical	Number of bedrooms	Predictor
Bathrooms	Numerical	Number of bathrooms	Predictor
Land Size_Cleaned	Numerical	Property land size	Predictor
Building_Area	Numerical	Unified building size (maximum of villa and house size fields)	Predictor
District_Price_Level	Numerical	Mean listing price within the property's district	Predictor
City	Categorical	Property location (city)	Predictor
Property_Type	Categorical	Type of property (e.g., land, villa, house)	Predictor
Price_Cleaned	Numerical	Property price (USD)	Target

Table 3. Final feature set used to train the Histogram Gradient Boosting model.

3.4 Data Cleaning and Preprocessing

Raw listings were transformed into a modelling-ready dataset through six preprocessing steps documented in the project material: (1) currency and numeric cleaning, which removed non-numeric characters from the price field and converted it into a standardized Price_Cleaned column; (2) text-based feature extraction, using text mining on listing descriptions and specifications to recover Land Size_Cleaned, Villa Size_Cleaned, and House Size_Cleaned where these were present only as free text; (3) geographical data parsing, which split the raw location string into village, district, and city fields and standardized inconsistent spellings; (4) categorical classification, which identified property type (villa, land, or house) and title type from keywords in the listing title and description; (5) data pruning and dimensionality reduction, which removed high-cardinality or redundant columns such as property code, listing URL, and raw description text once the relevant structured fields had been extracted; and (6) schema standardization, which brought fields such as bedrooms, bathrooms, and garages into a consistent format across all records.

At the modelling stage, the implemented pipeline applies two further preprocessing steps directly in code. Numeric features (Bedrooms, Bathrooms, Land Size_Cleaned, Building_Area, and District_Price_Level) pass through a SimpleImputer configured with strategy='median', which fills any remaining missing numeric values with the column median, followed by a RobustScaler, which standardizes each feature using statistics that are resistant to outliers. The two categorical features (City and Property_Type) are transformed using a OneHotEncoder configured with handle_unknown='ignore', so that any category not seen during training is safely ignored rather than causing an error at prediction time. A

ColumnTransformer combines the numeric and categorical pipelines into a single preprocessing step, and the target price is transformed using a TransformedTargetRegressor with np.log1p and its inverse np.expm1 (Section 3.6).

3.5 Outlier Treatment

The project's descriptive presentation material and its implemented code describe two related but not identical outlier-filtering strategies, and in keeping with the project's own accuracy requirements this paper reports the code-based method as authoritative. The presentation material describes filtering "the top 2%" of price outliers (elsewhere referred to as a "0.98 quantile" filter) as part of the improvement that raised the model's R^2 from 0.56 to 0.9373. The implemented walkthrough code, however, applies a two-sided percentile filter: records are retained only if their price falls between the 5th and 95th percentile of Price_Cleaned, and land size is separately capped at its 95th percentile. This is a broader trim than a single-sided top-2% cut, since it removes both extremely low and extremely high price outliers. This paper describes the implemented 5th–95th-percentile method as the outlier-treatment approach actually used to produce the results reported in Section 4, while noting that the presentation material frames the same general improvement — from an initial R^2 of 0.56 to a final R^2 of 0.9373 — using the "top 2%" / "0.98 quantile" description.

3.6 Feature Engineering

Two engineered features were central to the final model. First, Building_Area was constructed by taking the row-wise maximum of the Villa Size_Cleaned and House Size_Cleaned columns, giving a single unified measure of building size regardless of property type. Second, District_Price_Level was constructed as a location-intelligence feature: for each district, the mean Price_Cleaned of properties within that district was computed and mapped back onto every record in that district. This allows the model to learn location-related value without one-hot-encoding every individual district name, which would otherwise create a very large number of sparse columns for a district-level categorical variable.

The target variable, Price_Cleaned, was log-transformed using np.log1p before training and reversed using np.expm1 after prediction, via scikit-learn's TransformedTargetRegressor. Because Cambodian property prices are highly right-skewed — with a small number of very high-value "trophy" listings alongside a much larger number of typical-market listings — a raw price target would let a small number of extreme values dominate the model's error signal. The logarithmic transformation compresses the scale of these extreme values, allowing the model to focus on the price patterns that describe the bulk of the market rather than being disproportionately influenced by a handful of luxury outliers.

3.7 Machine Learning Model

The primary algorithm used in this project is the Histogram Gradient Boosting Regressor (HGBR), implemented via scikit-learn's HistGradientBoostingRegressor. HGBR is a tree-based ensemble method that builds an additive sequence of shallow decision trees, where each new tree is trained to correct the residual errors left by the trees built before it. Before splitting, HGBR bins each continuous feature into a fixed number of histogram bins, which substantially reduces the computational cost of finding the best split point at each node compared with evaluating every unique feature value. This binning strategy makes

HGBR well suited to the kind of large, mixed-type tabular data found in property listings: it can model non-linear relationships and feature interactions (for example, the way land size and district jointly influence price) more flexibly than a linear model, it trains efficiently even as the number of records grows, and — critically for this project's noisy, real-world Cambodian data — it can handle missing feature values natively during training, providing an additional layer of robustness beyond the explicit SimpleImputer step already applied in preprocessing.

3.8 Model Configuration

The final model configuration, taken directly from the implemented walkthrough code, is summarised in Table 4. No hyperparameters beyond those listed were used.

Parameter	Value
Algorithm	HistGradientBoostingRegressor (scikit-learn)
random_state	42
max_iter	500
learning_rate	0.03
Target transformation	np.log1p (forward), np.expm1 (inverse), via TransformedTargetRegressor
Train/test split	85% train / 15% test (test_size = 0.15)
Split random_state	42

Table 4. Model configuration used to train the final Histogram Gradient Boosting model.

3.9 Training and Testing

The 880-record filtered dataset (Section 3.2, Section 4.1) was split into an 85% training set and a 15% held-out test set using scikit-learn's train_test_split with test_size=0.15 and random_state=42. It should be noted that this 85/15 split differs slightly from the 90/10 split described conceptually in the project's presentation material; the implemented code uses the 85/15 ratio, and all results reported in Section 4 are based on this split. Holding back an unseen test set is essential for an honest evaluation: because the model never sees the test records during training, the metrics computed on this held-out set (Section 3.10) reflect how well the model is likely to generalize to new, previously unseen Cambodian property listings, rather than simply how well it has memorized the training data.

3.10 Evaluation Metrics

Model performance was assessed using four metrics, each capturing a different aspect of prediction quality.

Metric	Definition	Practical Interpretation
R^2 (coefficient of determination)	Proportion of the variance in price explained by the model.	Values closer to 1.0 indicate the model explains most of the variation in property prices.

Metric	Definition	Practical Interpretation
MAE (Mean Absolute Error)	Average absolute difference between predicted and actual price, in USD.	Gives the typical dollar-value size of a prediction error.
RMSE (Root Mean Squared Error)	Square root of the average squared prediction error.	Penalizes larger errors more heavily than MAE; not reported numerically in the supplied project evidence (Section 4.3).
MAPE (Mean Absolute Percentage Error)	Average absolute error expressed as a percentage of the actual price.	Expresses typical error as a relative percentage, comparable across price ranges.
Accuracy within ±20%	Share of test-set predictions whose absolute error is less than 20% of the actual price.	A practical, bank-appraisal-style tolerance band used to judge whether a prediction is “close enough” for real-world use.

Table 5. Evaluation metrics used to assess the Histogram Gradient Boosting model on the held-out test set.

4. Results

4.1 Dataset Results

The working dataset began with 1,001 initial listing records, all of which contained a valid Price_Cleaned value at the data-loading stage (Section 3.2). After the price- and land-size-based outlier filtering described in Section 3.5, the final modelling dataset contained 880 records and seven predictor features, plus the Price_Cleaned target. This final dataset was split into 748 training records (85%) and 132 test records (15%, test_size = 0.15). Two data-quality figures are reported directly from the project's exploratory analysis: approximately 38% of records were missing a building-area value before the Building_Area feature was engineered, and price itself was reported as approximately 1.6% missing at an earlier stage of analysis — a figure that, as noted in Section 3.2, is not fully consistent with the final loading code, which found no missing price values among the 1,001 initial records. On missing land-size values specifically, the project's own supporting slides disagree with one another: one slide reports approximately 17% missing land size, while a separate data-completeness chart in the same project reports 0% missing land size. Both figures are reported here rather than silently reconciled, since the underlying project evidence does not resolve the discrepancy.

4.2 Exploratory Data Analysis

Prior to preprocessing, listing prices were found to be highly right-skewed, with a reported median of approximately $175,000 against a maximum of approximately $6.5 million, reflecting a small number of very high-value listings alongside a much larger number of typical-market properties (Figure 1). This skew is the main motivation for the log-transformation of the target price described in Section 3.6.

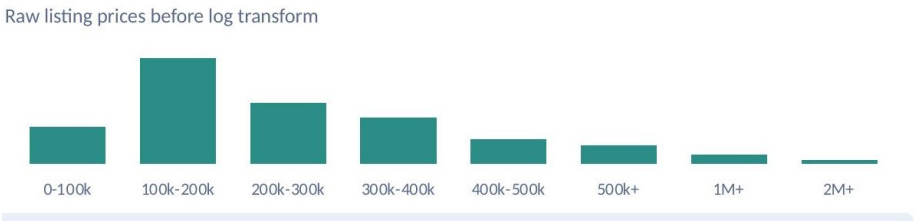

Figure 1. Distribution of raw Cambodian property listing prices prior to log transformation, showing a right-skewed distribution with a concentration of listings in the $100,000–$300,000 range.

A correlation analysis of the raw, categorically encoded feature set (Figure 2) shows that Bedrooms and Bathrooms are strongly correlated with one another (r = 0.93), as expected for similarly sized properties, and that Land Size_Cleaned and Villa Size_Cleaned are also strongly correlated (r = 0.94). Price_Cleaned shows a moderate positive correlation with Bathrooms (r = 0.21), Bedrooms (r = 0.19), and House Size_Cleaned (r = 0.25), a weak negative correlation with Land Size_Cleaned (r = −0.06) — likely reflecting the very wide range of land parcel sizes in the raw data — and weak negative correlations with District (r = −0.14) and City (r = −0.18) as encoded categorical values. These are simple linear (Pearson-style) correlations on the raw, unfiltered data and should not be interpreted as a ranking of feature importance for the trained HGBR model itself, since HGBR captures non-linear and interaction effects that a pairwise correlation coefficient cannot represent (see Section 4.6).

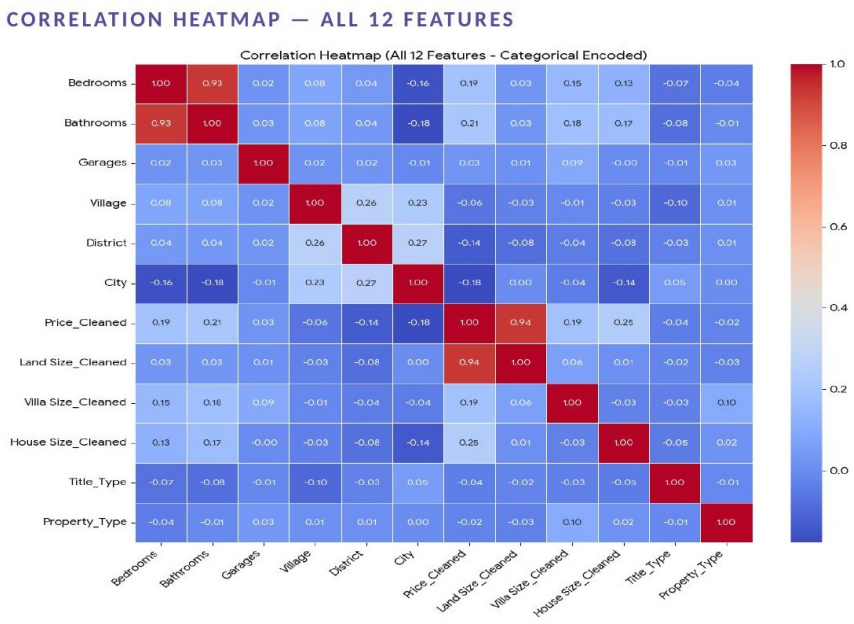

Figure 2. Correlation heatmap of raw property features (categorically encoded), showing pairwise linear relationships prior to feature engineering and outlier filtering.

The property listings in the dataset are mixed in type, with land listings forming the largest share by volume, followed by villas, with condominiums representing a smaller but notable share of the market (Figure 3). Median prices also varied by city, with Phnom Penh listings commanding a substantially higher median price than provincial markets such as Siem Reap, Sihanoukville, and Kampot, consistent with the capital-city price premium that motivated the District_Price_Level feature described in Section 3.6.

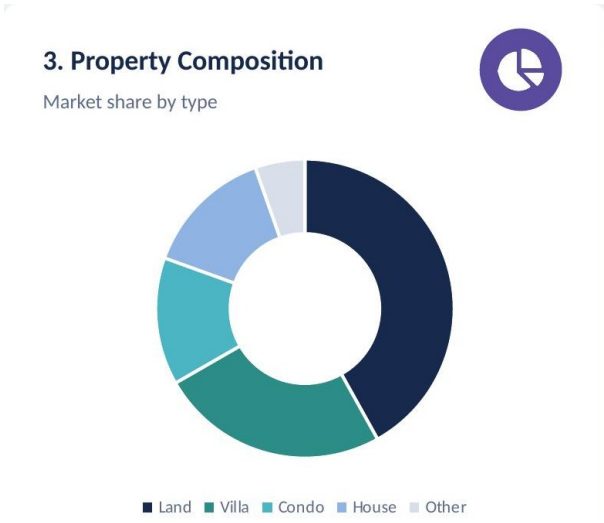

Figure 3. Composition of the property dataset by property type, showing land listings as the largest category by volume, followed by villas and condominiums.

4.3 Model Performance

On the 132-record held-out test set, the final HGBR model produced the results shown in Table 6. The model explains 93.73% of the variance in property price on unseen data, has a Mean Absolute Percentage Error of 6.01%, and places 96.97% of its predictions within ±20% of the true listing price. Mean Absolute Error is reported as $34,617.38, taken directly from the project's printed evaluation output. RMSE was not reported anywhere in the supplied project evidence and is therefore not stated here, in line with the project's own accuracy requirement not to report a value that was not actually produced by the implemented code.

Metric	HGBR Result
R^2	0.9373
MAPE	6.01%
Accuracy within ±20%	96.97%
MAE	$34,617.38
RMSE	Not reported in the supplied project evidence

Table 6. Histogram Gradient Boosting model performance on the 15% held-out test set.

4.4 Actual vs. Predicted Prices

Figure 4 plots predicted price against actual price for every property in the test set, together with a fitted trend line. If the model produced perfect predictions, every point would fall exactly on the 45-degree diagonal. In practice, the points cluster tightly around the diagonal across the full range of test-set prices shown (roughly $200,000 to $1,400,000+), with the fitted trend line closely tracking the diagonal. This visual pattern is consistent with the high R^2 and low relative error reported in Table 6, and shows that the model's accuracy is not concentrated only in the lower or middle part of the price range but extends across most of the observed price spectrum, with some widening of the prediction band at the highest prices shown, where fewer training examples are available.

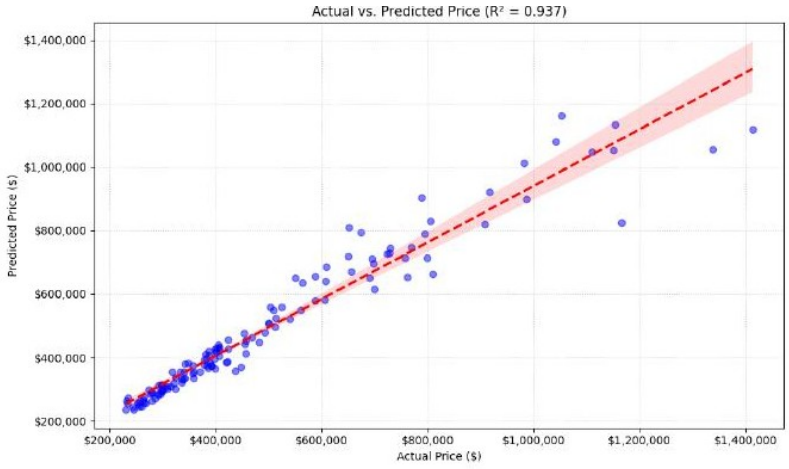

Figure 4. Actual vs. predicted property prices on the held-out test set (R^2 = 0.937), with a fitted trend line shown in red.

4.5 Error Analysis

The reported metrics and the actual-vs-predicted plot together suggest that prediction error is generally modest and broadly distributed rather than concentrated in a single price segment, but a small amount of additional deviation is visible at the upper end of the price range in Figure 4, where two points in particular sit further from the diagonal. This pattern is consistent with what would be expected given the project's outlier-filtering strategy (Section 3.5): because the training data was filtered to the 5th–95th percentile of price and land size, the model was trained primarily on "typical market" listings, and its accuracy on the small number of very high-value or unusually large properties that remain in the test set may be somewhat lower than its accuracy on the bulk of the market. The project's own EDA also flagged missing building-area values (approximately 38% of raw listings) and inconsistent land-size reporting as sources of noise; while the SimpleImputer and RobustScaler steps in the preprocessing pipeline (Section 3.4) are designed to reduce the impact of this noise, some residual effect on individual predictions cannot be ruled out from the evidence available. No breakdown of error by city, district, or property type was produced by the supplied code, so this paper does not claim a specific location- or type-based error pattern beyond what is directly supported by the overall test-set metrics in Table 6.

4.6 Feature Importance / Model Interpretation

The implemented walkthrough code does not compute a formal feature-importance or permutation-importance output for the trained HGBR model, and no such result is available in the supplied project evidence. Reporting a specific feature-importance ranking here would therefore not be supported by the project material. The correlation analysis presented in Section 4.2 (Figure 2) offers a partial, pre-modelling view of which raw features move together with price, but a Pearson-style correlation coefficient reflects only linear, pairwise association and cannot substitute for a proper model-based importance measure such as permutation importance or SHAP values, which would account for HGBR's non-linear structure and feature interactions. Computing and reporting such a measure is identified as a specific direction for future work in Section 6.

5. Discussion

5.1 Main Findings

The evaluation results reported in Section 4.3 indicate that a Histogram Gradient Boosting model, trained on a modest 880-record Cambodian property dataset, can explain the large majority (93.73%) of the variance in listing price and can place the substantial majority (96.97%) of its predictions within a practically useful ±20% band of the true price. A Mean Absolute Percentage Error of 6.01% is low in absolute terms and suggests that, at least for the "typical market" listings retained after outlier filtering, the model's predictions are reasonably close to the properties' actual listed prices. Taken together, these results provide affirmative evidence for the first two research questions posed in Section 1.6: several property features (captured through Bedrooms, Bathrooms, Land Size, Building Area, District Price Level, City, and Property Type) jointly explain most of the variation in price, and Histogram Gradient Boosting is capable of producing accurate property-price predictions from this feature set. The third research question — how well the model generalizes across different locations and property types — is only partially answered by the available evidence, since the supplied project material does not include a location- or type-stratified breakdown of test-set error (Section 4.5).

5.2 Comparison with Previous Studies

The HGBR result obtained here (R^2 = 0.9373, MAPE = 6.01%) sits at or above the benchmark figures referenced in the project's supporting literature for other tree-based ensemble methods applied to real-estate data, including an R^2 of approximately 0.93 reported for an XGBoost model and an R^2 above 0.90 reported for a Random Forest model (Table 1). This should be interpreted cautiously: the benchmark studies were conducted on different datasets, in different markets, with different feature sets and evaluation protocols, and the exact bibliographic details behind several of these benchmark figures could not be fully verified from the project material (Section 8). The comparison nonetheless suggests that HGBR's accuracy on this Cambodian dataset is broadly in line with — rather than markedly worse than — accuracy levels reported for comparable ensemble methods elsewhere, which is a reasonable basis for cautious optimism about the approach rather than a claim of outright superiority.

5.3 Why HGBR Performs Well

Several design choices likely contributed to the model's performance. First, HGBR's ability to model non-linear relationships and feature interactions is well suited to property pricing, where the effect of one feature (for example, land size) on price plausibly depends on the value of another feature (for example, district). Second, the ensemble structure of gradient boosting, in which successive trees correct the residual errors of previous trees, tends to produce more accurate predictions than a single decision tree while remaining computationally efficient through histogram-based binning. Third, the preprocessing pipeline — median imputation, robust scaling, and one-hot encoding — reduces the impact of missing values and outlying numeric values before they reach the model. Fourth, the engineered District_Price_Level feature gives the model a compact, informative summary of location value without the sparsity problems of one-hot-encoding every district. Finally, the log-transformation of the target

price (Section 3.6) prevents a small number of extreme high-value listings from dominating the training loss, allowing the model to fit the bulk of the market more precisely.

5.4 Cambodian Context

The results are specifically relevant to the Cambodian residential market in several respects. The dataset spans multiple cities and property types, and the District_Price_Level feature explicitly captures the substantial price differences observed between Phnom Penh and provincial markets (Section 4.2). The project's problem statement also highlights Cambodia-specific complexities — including missing or inconsistently reported building-area and land-size data, and the presence of extreme high-value listings — that a generic off-the-shelf valuation heuristic would struggle to accommodate, and that the outlier-filtering, imputation, and log-transformation steps described in Section 3 were specifically designed to address. At the same time, because the dataset and model evidence available for this paper does not include separate performance figures for title-type (Hard Title vs. Soft Title) or district-level accuracy breakdowns, any claim about how strongly legal title status or fine-grained locality drives Cambodian property value would go beyond what the supplied evidence directly supports.

5.5 Practical Applications

A validated model of this kind has several plausible practical applications for the groups identified in Section 1.7. For home buyers and sellers, the model could provide an independent, transparent reference price to inform negotiation. For real estate agents, it could support listing price recommendations and client advice with a data-backed estimate. For property investors, it could help compare potential value across different cities, districts, and property types. For banks and lenders, a fast, consistent estimate could support — though not replace — mortgage and collateral assessment processes that currently depend on manual appraisal. For government bodies and researchers, the underlying dataset and model outputs could contribute to broader housing-market analysis. These are presented as plausible applications consistent with the project's stated aims rather than as claims that have been operationally validated with real users or institutions.

5.6 Limitations

- Dataset size — the final modelling dataset of 880 records, drawn from 1,001 initial listings, is modest by machine-learning standards and may limit the model's ability to generalize to property types or locations that are underrepresented in the data.
- Listing vs. transaction prices — the dataset represents property listings (asking prices) rather than confirmed sale (transaction) prices; the model should be understood as predicting asking-price levels rather than final negotiated sale prices.
- Missing and inconsistent data — building-area values were missing for a substantial share of raw listings (approximately 38%), and the project's own materials disagree on the extent of missing land-size data (Section 4.1); while imputation and robust scaling mitigate this, some residual uncertainty from data quality cannot be fully ruled out.
- Geographic coverage — the dataset covers a subset of Cambodian cities and provinces rather than the country as a whole, so model accuracy in provinces or districts with little or no representation in the training data is untested.

- Outlier handling — the 5th–95th percentile price filter and 95th-percentile land-size cap (Section 3.5) exclude the most extreme listings from training, which likely improves accuracy on typical-market properties but means the model's accuracy on very high-value or unusually large “trophy” properties has not been directly demonstrated.
- Market timing — property prices change over time with broader economic and market conditions; the model reflects the listing data available at the time of collection and would require periodic retraining to remain current.
- No formal feature-importance analysis — as discussed in Section 4.6, the supplied code does not include a permutation-importance or SHAP-based analysis, so specific claims about which individual features matter most to the trained model cannot be made from the available evidence.

6. Conclusion and Future Work

Cambodia's real estate market has grown rapidly, but property pricing remains largely informal, relying on agent judgement and word-of-mouth comparison rather than a consistent, data-driven process. This creates inconsistent pricing, an information gap between buyers and sellers, and inefficiency in bank and mortgage appraisal processes. This paper presented a machine-learning-based Automated Valuation Model for Cambodian residential property, built around a Histogram Gradient Boosting Regressor trained on a cleaned and feature-engineered dataset of 880 property listings (drawn from 1,001 initial records) spanning multiple Cambodian cities and property types.

The methodology combined currency and text-based data cleaning, geographic parsing, a unified Building_Area feature, a district-level price indicator, median imputation, robust scaling, one-hot encoding, and a log-transformed target within a single scikit-learn pipeline, trained with an 85/15 train-test split. On the held-out test set, the model achieved an R^2 of 0.9373, an MAE of $34,617.38, a MAPE of 6.01%, and placed 96.97% of predictions within ±20% of the true price. These results indicate that Histogram Gradient Boosting can model the non-linear, location-dependent relationships that shape Cambodian property prices with an accuracy broadly comparable to ensemble-method benchmarks reported in other real-estate markets, directly answering the study's first two research questions in the affirmative, while the third — generalization across locations and property types — remains only partially evidenced by the available material.

The practical significance of these results lies in the possibility of a faster, more consistent, and more transparent reference price for buyers, sellers, agents, investors, and lenders operating in a market that currently depends heavily on informal, relationship-based valuation. At the same time, the model's reliance on asking-price listings rather than confirmed transaction prices, its modest dataset size, and its untested performance on extreme high-value properties mean that it should be understood as a decision-support reference rather than a substitute for professional appraisal.

Future Work

- Expand the dataset with a larger volume of Cambodian property listings and, where possible, actual transaction-price data rather than asking prices.
- Extend geographic coverage to additional provinces and cities beyond those represented in the current dataset.
- Incorporate additional property attributes, such as legal title type, property age, and amenities, where reliably available.
- Add geographic or geospatial features (for example, distance to city centre or major roads) beyond the current district-level price indicator.
- Conduct systematic hyperparameter optimization for the HistGradientBoostingRegressor rather than relying on the current fixed configuration.
- Directly compare HGBR against XGBoost and Random Forest on the same Cambodian dataset, rather than relying on benchmark figures drawn from other markets.
- Apply explainable-AI techniques such as SHAP values to produce a rigorous, model-based feature-importance analysis.

- Establish a continuous retraining process so the model remains current as market conditions change.
- Explore deployment as a web-based Automated Valuation Model with a simple user interface for price estimation.
- Investigate integration with banking mortgage and collateral-assessment workflows, subject to appropriate validation and oversight.

Acknowledgment

The authors gratefully acknowledge Norton University and the Graduate School for the opportunity to undertake this Machine Learning course project. Sincere thanks are extended to the course lecturer, Sek Socheat, for guidance throughout the project and for feedback on the journal-style paper format. The authors also acknowledge the contributions of all Group 10 members — Oum Vuthy (model development), Chan Sovychin (data collection), Chan Reaksmey (data preparation), and Chhorn Menghong (evaluation and reporting) — whose combined efforts in data collection, preprocessing, model development, evaluation, and reporting made this study possible.

References

References are presented in IEEE style with numbered in-text citations. Several entries are given in the project's supporting slide material only as an author name and year, without a journal title, volume, page range, or DOI. Rather than fabricate this missing bibliographic information, such entries are reproduced here with the identifying details available in the source material and are explicitly flagged as requiring verification before formal submission or publication.

[1] Al Marzooqi and A. Redouane, "[Title not specified in source material] — XGBoost-based property price prediction study," 2025. Citation incomplete in the project's supporting material (journal, volume, and DOI not provided) — requires verification before formal submission.

[2] S. Sevgen and H. Tanrivermiş, "[Title not specified in source material] — Random Forest property valuation study (Istanbul context)," 2024. Citation incomplete in the project's supporting material (journal, volume, and DOI not provided) — requires verification before formal submission.

[3] A. Zilli et al., "[Title not specified in source material] — Artificial Neural Network vs. traditional Hedonic Pricing Model comparison," 2024. Citation incomplete in the project's supporting material (journal, volume, and DOI not provided) — requires verification before formal submission.

[4] "Standalone Decision Tree benchmark for property price prediction (R^2 = 0.385)," referenced in project material without an author or publication citation — requires verification before formal submission.

[5] Knight Frank, "[Report title not specified in source material] — 2025 real-estate market report referenced for a 'flight to quality' and market-stabilisation finding," 2025. Citation incomplete in the project's supporting material — requires verification before formal submission.

[6] S. Por and Sario, "[Study title not specified in source material] — survey on consumer priorities in property purchase decisions (legal title as top priority, mean = 4.38)," 2025. Citation incomplete in the project's supporting material — requires verification before formal submission.

[7] Realestate.com.kh, "[Survey/report title not specified in source material] — 2025 buyer-intent survey referenced in project discussion," 2025. Citation incomplete in the project's supporting material (specific report title and URL not provided) — requires verification before formal submission.

[8] Ja'afar and Mohamad, "[Title not specified in source material] — study on the suitability of tree-based machine-learning models for unpredictable/noisy data," 2021. Citation incomplete in the project's supporting material (journal, volume, and DOI not provided) — requires verification before formal submission.

[9] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[10] Group 10 (O. Vuthy, C. Sovychin, C. Reaksmey, C. Menghong), "Property Value Prediction Using Machine Learning, Cambodia" and "Property Valuation in Cambodia," Machine Learning course project material, Norton University, 2026. [Primary project source material for this paper.]

[11] Group 10 (O. Vuthy, C. Sovychin, C. Reaksmey, C. Menghong), "Code Walkthrough Script: Property Value Prediction Model — HistGradientBoostingRegressor Pipeline, Cambodia Property Data," Machine Learning

course project material, Norton University, 2026. [Primary source for the implemented preprocessing pipeline, model configuration, and evaluation results reported in this paper.]