# **ISY-5001 Intelligent Reasoning Systems**



**Project Name: Property AI Assistant**

**Submission Type: Final Report**

| **Student Name** | **Student ID** |
| --- | --- |
| Cheng Ziming | A0330176L |
| Hong Zhuoyuan | A0329586L |
| Jia XingJie | A0139137A |
| Thun Zhen Hong | A0331639B |
| Wang Yaochen | A0241873B |

**Content Page**

| **Table of Contents** | **Page** | |
| --- | --- | --- |
| **1. Introduction** | **3** | |
| **2. Project Background & Market Context** | **3** | |
| **3. Literature Review & Market Research** | **5** | |
| **4. Project Scope** | **7** | |
| **5. Data Collection and Preparation** | **8** | |
| **6. System Design** | **12** | |
| **7. Implementation** | **21** | |
| **8. Results & Progress** | **23** | |
| **9. Challenges & Roadblocks** | **24** | |
| **10. Future Work** | **25** | |
| **11. Conclusion** | | **25** | |
| **12. References** | | **26** | |
|  | |  | |

# Introduction

Singapore’s residential property market is characterized by high financial stakes, tight regulatory controls, and rapidly evolving buyer preferences. Property decisions, whether upgrading, downgrading, or investing, require navigating a complex mix of affordability rules, market timing, and lifestyle considerations. To address this challenge, this project proposes an Intelligent Property Reasoning System that integrates data-driven price estimation, personalized recommendation modeling, and LLM-based spatial reasoning to support informed, end-to-end decision-making for Singapore homebuyers and sellers.

The system combines multiple analytical layers to deliver comprehensive insights. The HDB Property Price Estimator, powered by an XGBoost regression model, predicts entry and exit prices based on structural, locational, and temporal attributes. The Property Recommendation Model, built upon a LightGBM Ranker trained under the LambdaMART framework, ranks candidate flats using user preference vectors and similarity mapping functions. A lightweight LLM agent layer further enhances interpretability and user interactivity, performing tasks such as route visualization and affordability validation through dynamic rule checking against financial benchmarks including TDSR, MSR, BSD, and ABSD.

By fusing government-verified datasets from HDB, URA, and OneMap with user preference matrices and geospatial intelligence, the system aims to reduce uncertainty, accelerate decision-making, and enhance transparency in property planning. It provides actionable intelligence for a wide range of stakeholders including first-time buyers, investors, upgraders, downgraders, agents, and mortgage advisors, enabling decisions that are both financially sound and aligned with lifestyle and locational priorities.

1. Project Background & Market Context

Singapore’s real estate market in 2025 shows a measured cool-down rather than a contraction. Public housing resale prices moderated to +0.9% Quarter-on-Quarter (QoQ) in Q2 2025, with the Resale Price Index at 202.9. This signals a softer but still positive trend after several years of rapid gains (HDB, 2025).

Private residential prices rose 1.0% QoQ over the same period, with a differentiated pattern across segments: Core Central Region (CCR) +3.0%, Rest of Central Region (RCR) −1.1%, Outside Central Region (OCR) +1.1%. This shows the growing importance of location factor in price point (URA, 2025). On the supply side, 10,209 flats were launched in July 2025 through the combined BTO and SBF exercises, demonstrating the government’s resolve in controlling the public supply of housing and ensuring affordability and choice (HDB, 2025). These new launches operate under the Standard/Plus/Prime flat classification introduced in October 2024, which will greatly affect the property market in the future. (HDB, 2024; MND, 2025).

The future outlook for the affordability of Singapore housing also looks to be good. The Total Debt Servicing Ratio (TDSR) remains at 55%, ensuring the public does not over-commit to houses they cannot afford the monthly payments (MAS, 2021). For HDB buyers, the loan-to-value (LTV) limit was reduced to 75% with effect from 20 August 2024, raising upfront cash/CPF requirements and affecting decisions for potential house flippers and upgraders. (HDB, 2024; MND, 2024).

Additional Policies that limit the rapid growth of properties are also in place. Buyer’s Stamp Duty (BSD) retains a 6% top marginal rate on higher-value residential purchases (IRAS, 2023, 2025), while Additional Buyer’s Stamp Duty (ABSD) of up to 60% limits the buying intention of people who already own multiple properties. However, married Singapore-citizen couples may obtain ABSD remission if their first property is sold within six months (IRAS, 2025). For private properties, Seller’s Stamp Duty (SSD) was tightened from 4 July 2025, extending the holding period to four years and raising rates at every tier, thereby elevating the cost of premature exits (IRAS, 2025; MAS, 2025). These policy measures make entry and exit choices both policy-aware and time-sensitive for the public.

Mainstream property solutions in the market use data aggregation and other calculative methods. 99.co/SRX provides an Automated Valuation Model that is useful for price estimates but does not consider policy-aware progression planning (99.co/SRX, 2025). PropertyGuru offers valuation calculators and trend tools that assist in benchmarking, but also do not consider ABSD/SSD constraints (PropertyGuru, 2024, 2025). EdgeProp’s Edge Fair Value and Deal Checker help in pricing (EdgeProp, 2025). Ohmyhome (HomerAI) uses human-focused appointment setting rather than neutral rules-based optimisation (Ohmyhome, 2025). MOGUL.sg (MAIA) showcases possible AI-based search, yet its constraint parameters remain limited (MOGUL.sg, 2025).

This leaves a clear market gap for a system that unifies predictive accuracy with rule-aware reasoning. The proposed assistant addresses that gap by integrating:

(i) A robust AVM with uncertainty bands.

(ii) Hard policy/finance constraints (TDSR, LTV, ABSD/BSD/SSD).

(iii) Sequence/timing optimisation for sell-then-buy versus buy-then-sell.

(iv) Amenity-adjusted recommendations using map routing and points of interest (schools, transport, parks, retail).

(v) Grants Helper that computes net-effective prices using EHG, CPF Housing Grant, and PHG (up to S$230k for eligible first-timer families).

(vi) HFE Letter alignment for cradle-to-decision eligibility and loan/grant checks.

(vii) LLM Guidance that transparently flags Standard/Plus/Prime resale restrictions and subsidy recovery implications for long-run planning.

In summary, the opportunity is to elevate today’s solutions into a AI copilot/Assistant that is accurate, comprehensive, personalised and future planning to Singapore residence looking at their property needs.

1. Literature Review & Market Research

The study of property valuation has traditionally centred on the hedonic pricing model, which explains housing values through characteristics like structural, locational, and neighbourhood traits. While this framework remains foundational in housing economics, it often struggles to capture complex, non-linear interactions and spatial dependencies (Khoshnoud & Sirmans, 2023). To overcome these limitations, machine learning methods have gained prominence in predicting housing prices. Among them, gradient-boosted decision trees, particularly XGBoost, have been recognised as strong baselines for house price prediction. These models handle high-dimensional data effectively, account for non-linear relationships, and offer interpretability through feature importance measures, making them well-suited for real estate applications (Sharma, Harsora, & Ogunleye, 2024).

Recent advances further highlight the importance of incorporating relational and spatial features into predictive frameworks. Graph-based models, such as graph neural networks (GNNs), enable the encoding of relationships between properties, points of interest, and neighbourhood amenities. This approach not only enhances predictive accuracy but also improves the explainability of recommendations by reflecting how spatial connectivity influences housing values (Karamanou, Brimos, Kalampokis & Tarabanis, 2024). For instance, proximity to transport infrastructure, schools, and commercial centres can be naturally integrated within a graph structure, offering a richer representation than conventional tabular inputs.

Taken together, these studies demonstrate an evolution from attribute-focused valuation toward integrated and data-driven approaches. Hedonic pricing remains theoretically important, but its predictive capacity is substantially augmented when combined with advanced learning methods. Machine learning and graph-based reasoning extend the scope of traditional models, aligning with the broader push for intelligent decision-support systems in real estate. In markets such as Singapore, where regulatory constraints on affordability and taxes significantly affect decisions, these approaches provide a foundation for more accurate, transparent, and user-centred property recommendations for consumers.

Explainability is increasingly emphasised in high-stakes financial domains, as users demand transparency in how entry and exit price predictions and recommendations are derived (Doshi-Velez & Kim, 2017). Building on these strands, the proposed system integrates LLM agents, domain-specific tools, and multimodal databases into a layered architecture. This design aligns with recent calls for modular, explainable, and user-centred AI systems that balance predictive accuracy with regulatory compliance and user trust (Gunning & Aha, 2019).

Property portals in Singapore, including PropertyGuru, 99.co, and SRX, excel at aggregating listings and past transactions, offering users transparency and convenient search features. These tools are strong in data coverage but primarily descriptive, focusing on comparisons of price, size, and location. Personalised guidance for long-term planning is minimal.

In addition, current portals generally lack tools that simulate upgrade or downgrade scenarios across time horizons, taking into account regulatory constraints such as Total Debt Servicing Ratio (TDSR), Mortgage Servicing Ratio (MSR), Buyer’s Stamp Duty (BSD), and Additional Buyer’s Stamp Duty (ABSD). Optimisation engines that align affordability, market timing, and lifestyle needs (e.g., proximity to schools or transport) are notably absent.

A graph with text and numbers

AI-generated content may be incorrect.

Figure 1: Matrix generated by ChatGPT prompt on competitive landscape

This leaves an important market gap. Few solutions combine comprehensive data with personalised, rule-aware optimisation. The proposed system is designed to fill this space by predicting entry and exit prices, simulating upgrade or downgrade scenarios, and aligning recommendations with both financial feasibility and lifestyle needs. This enables faster and more confident property decisions.

1. Project Scope

This project develops an intelligent reasoning system designed to assist Singapore homebuyers and sellers in making data-driven decisions when upgrading or downgrading their properties. The system integrates three functional pillars — predictive modeling, rule-based affordability evaluation, and personalized recommendations — into a single cohesive platform rather than isolated tools. In its initial phase, the focus lies on constructing a transparent price-prediction module for both HDB and private properties, incorporating interpretable model outputs and uncertainty ranges. Alongside this, an affordability engine implements core financial and regulatory checks such as TDSR, MSR, BSD, and ABSD to ensure realistic and policy-compliant affordability assessments. Complementing these is an amenity-sensitive recommender that aligns housing suggestions with lifestyle preferences, considering factors such as proximity to schools, transport connectivity, parks, and urban convenience.

A conversational interface supports natural-language interaction, enabling users to query the system intuitively (e.g., “Find a 4-room flat near Punggol MRT under S$700K”). While the scope covers valuation, recommendation, and rule-based reasoning, advanced features such as legal contract automation or loan approval workflows are reserved for future development phases.

# Data Collection & Preparation

**5.1 Data Sources**

5.1.1 HDB Resale Transactions

The Housing & Development Board (HDB) Resale Flat Prices dataset is sourced from the official Singapore government open data portal, data.gov.sg. It contains transaction-level records of all HDB resale flats from 2017 to the present, updated periodically. Each record provides detailed information on the transacted unit, including:

* Transaction month, town, flat type, block, street name, and storey range
* Floor area (sqm), flat model, lease commencement date, and remaining lease
* Resale price (SGD).

This dataset serves as the foundation for our price estimation and feature engineering pipeline, offering a structured and government-verified view of Singapore’s public housing market.

5.1.2 URA Private Residential Transactions

The Urban Redevelopment Authority (URA) publishes data on private residential property transactions through eservice.ura.gov.sg. Our project utilizes URA records from 2020 onwards, covering both new sales and resale transactions.

Each record includes the following attributes:

* Project name, transacted price, floor area (sqft/sqm), and unit price ($PSF)
* Transaction date, type of sale (new/resale), and type of area (e.g., strata)
* Property type (condominium/apartment), tenure (leasehold/freehold), and market segment (CCR/RCR/OCR)
* Floor level

These attributes enable comparative modeling of private residential pricing trends and allow cross-segment analysis between the HDB and private housing markets.

5.1.3 Geospatial and Amenity Data

To capture locational context, we integrate a range of geospatial datasets from data.gov.sg and related open data repositories. These include:

* Education: Locations of primary, secondary, and tertiary schools
* Land Use: Zoning and planning boundaries
* Recreation and Environment: Parks and green spaces
* Commerce and Transport: Commercial facilities, MRT stations, and bus interchanges

By linking these datasets spatially to property locations, we enrich each record with contextual and lifestyle-related features, allowing the system to model real-world housing preferences—such as proximity to schools for families, access to public transport for commuters, and nearby amenities for lifestyle convenience.

5.1.4 OneMap API Integration

To enhance geospatial precision and accessibility modeling, we integrate the OneMap API, the national mapping platform developed by the Singapore Land Authority (SLA).

The API provides services for:

* Geocoding and reverse geocoding (postal code ↔ coordinates)
* Route and travel-time computation (e.g., walking or transit distance)
* Points of Interest (POI) retrieval for schools, healthcare, parks, and transport facilities

Through OneMap, the system dynamically computes commute times, walkability scores, and accessibility indices, enriching both the recommendation engine and the price estimator with real-time spatial intelligence.

**5.2 User Matrix Data**

In the recommendation model, the User Matrix represents each homebuyer’s preference profile across various housing features, rather than explicit numeric ratings. Each user’s vector captures how strongly they value attributes such as location proximity, flat size, building age, or lease duration. These values are later compared with the actual characteristics of candidate properties through the system’s feature mapping functions, which transform each user–property pair into normalized similarity scores ranging from 0 to 1.

User preference data is derived from both the mocked up Singapore Homebuyer Profiles and Condominium User Profiles datasets. Together, these matrices form the foundation for learning personalized ranking and relevance estimation, allowing the system to reflect nuanced user needs such as affordability sensitivity, lifestyle priorities, and location trade-offs.

| **Feature Name** | **Type** | **Description** | **Value Range / Encoding** |
| --- | --- | --- | --- |
| budget\_range | Numeric | User’s maximum purchase budget, used for filtering candidate properties. | Continuous (SGD) |
| preferred\_town | Categorical | Primary HDB town or region of interest. | Encoded via planning area index |
| property\_type\_preference | Categorical | User’s preferred property segment. | {HDB, Condo, Exec Condo} |
| flat\_type\_importance | Ordinal | Importance level of specific flat sizes (e.g., 3-room, 4-room). | [0–1] normalized |
| floor\_area\_priority | Numeric | Relative importance assigned to larger floor area. | [0–1] normalized weight |
| building\_age\_tolerance | Numeric | Acceptable building age range; lower values indicate newer-flat bias. | [0–1] |
| remaining\_lease\_weight | Numeric | Sensitivity to lease duration in property evaluation. | [0–1] |
| proximity\_to\_transport | Numeric | Preference for accessibility to MRT/bus stations. | [0–1] |
| proximity\_to\_schools | Numeric | Indicates family-based need for nearby schools. | [0–1] |
| green\_space\_preference | Numeric | Interest in parks and recreational amenities nearby. | [0–1] |
| price\_sensitivity | Numeric | Reflects how price variations affect the user’s utility. | [0–1] |
| amenity\_density\_tolerance | Numeric | Comfort with dense urban environments (e.g., malls, offices nearby). | [0–1] |

*Table 1: User Matrix Features*

**5.3 Spatial Feature Engineering**

The feature engineering phase entailed comprehensive data integration and spatial analysis to enhance the realism and contextual relevance of the housing dataset. Supplementary datasets were sourced from various public repositories, encompassing information on educational institutions, transportation networks, shopping malls, parks, and food centers. Recognizing that proximity to such amenities significantly influences housing preferences, the objective was to derive quantitative geographic features that effectively represent each property’s surrounding environment.

Following the acquisition of the raw datasets, systematic procedures involving data cleaning, geocoding, and coordinate alignment were undertaken to ensure consistency with the existing housing records. Subsequently, spatial analysis techniques were applied to evaluate each property’s accessibility to multiple categories of facilities. These computations yielded a structured set of geo-features encapsulating the convenience and livability of each location. The process demanded rigorous validation of heterogeneous datasets obtained from diverse sources to maintain data integrity and analytical precision. The list of features generated through this process is presented below.

| **Field Name** | **Description** |
| --- | --- |
| **Plan** | Planning subzone classification where the property is located, referring to one of Singapore’s 55 planning subzones. |
| **HWKR\_500M** | Number of hawker centres accessible within 500 meters. |
| **MALL\_500M** | Number of shopping malls accessible within 500 meters. |
| **PK\_500M\_IN** | Indicator of whether a large park is accessible within 500 meters (1 = yes, 0 = no). |
| **GP\_SCH\_1K** | Number of good primary schools accessible within 1000 meters. |
| **GP\_SCH\_2K** | Number of good primary schools accessible within 2000 meters. |
| **bus\_200\_route** | Number of bus routes accessible within 200 meters. |
| **bus\_500\_route** | Number of bus routes accessible within 500 meters. |
| **mrt\_200\_line** | Number of MRT lines accessible within 200 meters. |
| **mrt\_500\_line** | Number of MRT lines accessible within 500 meters. |

*Table 2: Processed Geo Spatial Features*

**5.4 Temporal and Ordinal Feature Encoding**

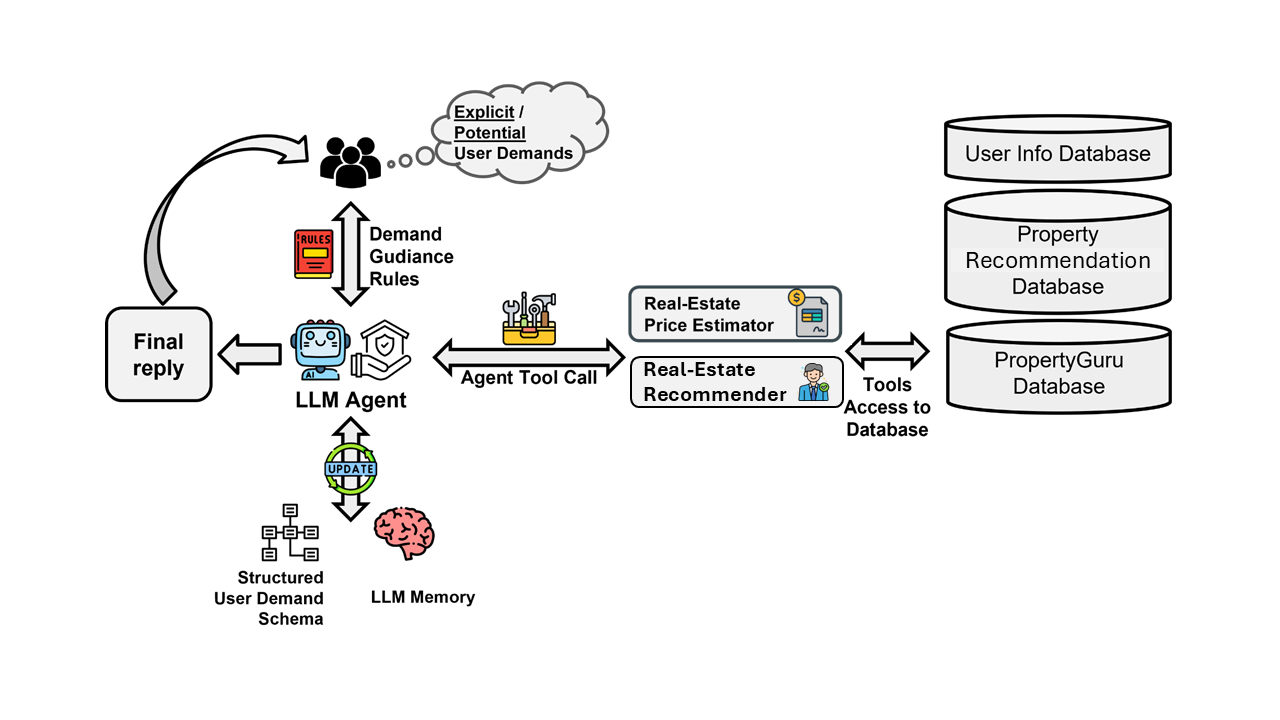
To effectively capture both macroeconomic and structural patterns in housing transactions, the temporal and ordinal variables were encoded using interpretable numerical transformations. The transaction month was decomposed into two components like “year” and “month\_num” to allow the model to learn long-term market appreciation trends and short-term seasonal effects. Meanwhile, the “storey\_range” attribute (e.g., “01 TO 03”, “04 TO 06”, “07 TO 09”) was converted into an ordered integer scale to represent the implicit vertical hierarchy of HDB buildings, ensuring the model recognizes that higher floors are typically more desirable and command higher resale prices.This encoding preserved both monotonic order and interpretability, which are crucial for tree-based models such as XGBoost.

Categorical housing features were encoded to retain class distinction while avoiding artificial ordinality. The flat\_type variable, representing different apartment classes (e.g., 2-Room, 3-Room, 4-Room, 5-Room, Executive, Multi-Generation), was one-hot encoded to generate binary indicator columns for each flat type. This approach preserved model transparency, allowing feature importance plots to directly reflect the contribution of each category without imposing a false numerical ranking among apartment sizes.

| **Feature Name** | **Source Attribute** | **Encoding Type** | **Description / Rationale** |
| --- | --- | --- | --- |
| year | month | Temporal decomposition | Extracted year component to capture long-term price trends. |
| month\_num | month | Cyclic temporal encoding | Numeric month index (1–12) to model seasonal variation. |
| storey\_level | storey\_range | Ordinal mapping | Encodes floor range as ordered integers reflecting height preference. |
| flat\_type\_2R, flat\_type\_3R, … | flat\_type | One-hot encoding | Creates separate binary columns for each flat type category. |
| region\_code | town | Label encoding | Assigns numeric region indices for planning areas. |

*Table 3: Processed Feature Encoding*

# System design



The system adopts a three-layer architecture comprising the Interaction Layer, Tool Layer, and Database Layer, each serving a distinct yet interdependent role in the reasoning workflow. At the core is the LLM Agent, which interfaces directly with users, interprets natural-language inputs, and orchestrates decision-making across the system. The Tool Layer operationalizes these decisions by invoking specialized modules such as the price estimator, recommendation engine, and rule-based calculators to execute targeted analytical or computational tasks. Beneath these components, the Database Layer provides structured access to housing, geospatial, and regulatory datasets, ensuring that every inference or recommendation is grounded in validated and contextually rich information.

**Interaction Layer**

**6.1 LLM agent design & functions**

6.1.1. Route Layout Function

To extend the system’s spatial functionality and improve user interaction, I developed a route layout function based on the Singapore OneMap API. This module enables users to specify an origin and a destination, upon which the system automatically computes and visualizes the optimal commuting route on an interactive map. The implementation combines routing algorithms and mapping visualization to provide users with an intuitive understanding of spatial connectivity between different locations.

By linking the recommendation results with real-world navigation, this function bridges analytical modeling with practical application. It not only supports users in evaluating housing choices from a spatial perspective but also enhances the overall usability and completeness of the Property AI Assistant platform.

6.1.2. LLM Agents

During user interaction, the LLM Agent serves as the central reasoning engine, orchestrating communication between the user, the analytical models, and the underlying databases. It continuously captures and structures user intents through a Structured User Demand Schema — a JSON-based representation that records parameters, constraints, confidence levels, and contextual dependencies. This schema evolves dynamically as the conversation progresses, while the agent’s long-term memory preserves dialogue history to maintain context across multiple interactions.

A rule-based controller guides the agent’s reasoning process, determining whether sufficient and unambiguous information has been gathered before proceeding. Once user requirements are clearly defined, the agent consolidates the inputs and issues a confirmation for validation. If ambiguities remain, it triggers an iterative refinement loop until the demand achieves mutual clarity and completeness.

When advancing to the searching, evaluation, or recommendation phase, the LLM Agent autonomously plans and executes tool usage. Each analytical module (Property Price Estimator and Property Recommendation Engine) is implemented as a callable tool within the system. The agent decides which tools to invoke, prepares the appropriate structured inputs, and calls them as part of its internal reasoning plan. After tool execution, the agent synthesizes and interprets the outputs, ensuring that the responses are accurate and contextually aligned with the user’s objectives.

**Tool Layer**

## **Property Price Estimator Model**

The HDB Property Price Estimator is a predictive system designed to estimate resale prices of Singapore HDB flats based on key structural, locational, and temporal factors. The system operates through a sequential workflow encompassing **data preprocessing and transformation, model training, validation, and temporal forecasting for inference**. The model was implemented using the XGBoost regression framework and trained on transaction records of HDB resale flats containing postal information, extracted and cleaned from open-source datasets.

6.2.1. Data Preprocessing and Transformation

Raw HDB resale transactions sourced from data.gov.sg are first validated for completeness and formatting integrity. Non-predictive identifiers such as town, street name, block number, and flat\_model are excluded to reduce dimensional noise and avoid redundancy, as these attributes are already captured implicitly through postal code and lease commencement year.

Temporal attributes such as month are decomposed into year and month\_num features to enable the model to distinguish between long-term macroeconomic appreciation trends (captured by year) and short-term market seasonality (captured by month\_num). This decomposition preserves time-series granularity, allowing the estimator to generalize across multiple years of resale cycles.

Lease-related attributes are also carefully restructured. The remaining\_lease field that is originally stored as a textual description (e.g., “62 years 3 months”) is programmatically converted into a continuous numeric value in months, denoted as remaining\_lease\_months. This transformation enables the regression model to learn the non-linear depreciation effects associated with shorter leases while maintaining interpretability for downstream forecasting.

Through these preprocessing mechanisms, it guarantees that every record follows a unified schema with consistent datatypes and chronological ordering. This foundation ensures that subsequent modeling stages can operate seamlessly on well-formed data without ad-hoc handling of missing or incorrect values.

6.2.2. Model training

Subsequently, the processed dataset was partitioned into 80% training and 20% testing subsets to preserve representative sampling across different towns, flat types, and time periods. This split ensures that the model generalizes well to unseen data while avoiding overfitting to localized temporal patterns.

XGBoost, a gradient-boosting framework based on decision trees, was selected for its ability to capture non-linear feature interactions, handle heterogeneous data types, and maintain robustness in the presence of missing values or outlier transactions. The model was configured with tuned hyperparameters to balance learning speed and generalization:

* **n\_estimators** = 500 ensures sufficient boosting rounds for convergence without overfitting.
* **learning\_rate** = 0.05 provides a moderate shrinkage rate, promoting gradual optimization and preventing instability in early boosting iterations.
* **max\_depth** = 6 constrains individual tree complexity, encouraging the model to learn intermediate-level feature interactions (e.g., between remaining\_lease\_months and storey\_level) rather than highly specific patterns that could harm generalization.
* **objective** = "reg:squarederror" minimizes the Mean Squared Error (MSE), directly aligning with continuous price prediction objectives

During training, the algorithm constructs decision trees sequentially, with each tree correcting the residuals of the previous ensemble. This process effectively captures the marginal contribution of each housing feature to price prediction — for instance, quantifying how an additional floor level or a longer remaining lease affects the estimated resale value. Feature importance metrics derived from the trained model provide insights into the dominant price determinants, typically highlighting floor area, location (postal code), and remaining lease duration as key drivers.

6.2.3. Model Validation

The validation and evaluation phase forms the empirical backbone of the HDB Price Estimator system, ensuring that the trained model demonstrates both predictive accuracy and generalization stability. Following model training, the system evaluates performance on an unseen 20% test subset using three key regression metrics like R², RMSE, and MSE which are computed directly from the predicted and actual resale prices.

The Coefficient of Determination (R²) quantifies the proportion of variance in actual resale prices that can be explained by the model’s predictions. A high R² value indicates that the model captures meaningful relationships among housing attributes such as floor area, storey level, and remaining lease duration. The Root Mean Squared Error (RMSE) measures the average magnitude of prediction deviations, expressed in the same units as the target variable (Singapore dollars). This provides an intuitive sense of how far, on average, predicted prices deviate from true transaction values. The Mean Squared Error (MSE) complements this by penalizing larger deviations more heavily, emphasizing stability and outlier sensitivity within the dataset.

Upon completion of validation, the trained XGBoost model incorporate all feature encodings and learned parameters that is serialized as hdb\_model.json for efficient reuse and deployment within the inference pipeline. This serialized artifact encapsulates both the learned decision ensemble and its associated hyperparameters, ensuring reproducibility and consistent inference performance across deployment environments.

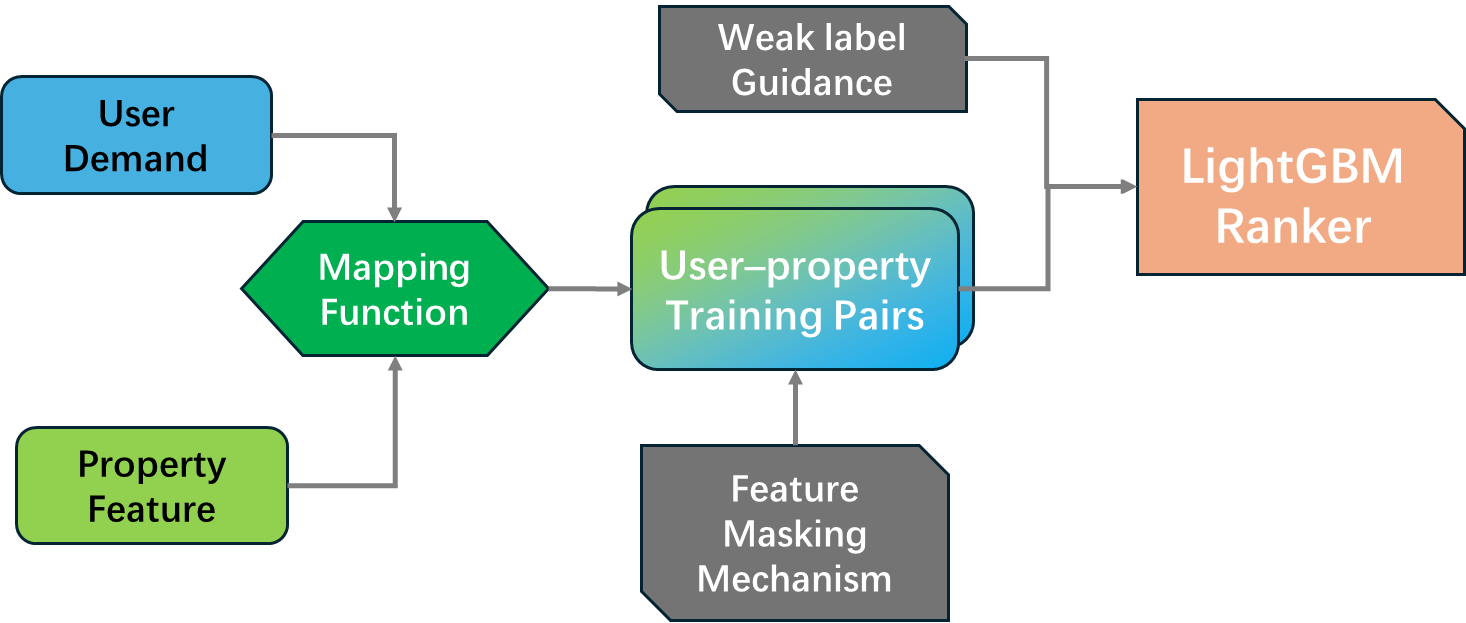
6.2.4. Temporal Forecasting for Inference

The inference and temporal forecasting module applies the trained XGBoost model to generate both current and forward-looking price estimates. During inference, user-specified inputs — including postal code, flat type, storey range, floor area, lease information, and transaction month — are processed through the same preprocessing and encoding pipeline used during training to ensure full consistency. The standardized input is then fed into the serialized model (hdb\_model.json), which outputs a predicted resale price for the given property configuration.

To forecast future price movements, the system employs a temporal simulation loop that incrementally advances the month and updates the remaining\_lease\_months at each step. The model generates a new prediction for each period, forming a time series of projected prices that reflects expected value changes as the property ages.

Predicted values are visualized alongside historical transaction averages to highlight model alignment with past trends. A ±RMSE confidence band is added to indicate the typical prediction uncertainty derived from validation. The resulting output provides both a point estimate and a forecast trajectory, enabling interpretable, data-driven insights into the evolution of HDB resale values over time.

## **Property Recommendation Model**

The implementation process of the model is illustrated in the figure below:

The recommendation model operates through five main stages: **feature mapping, sample generation, model training & validation, and inference**. User preferences are first transformed into quantitative similarity measures that compare each property’s attributes with the user’s stated needs. During sample generation, a feature masking mechanism randomly hides portions of the user-driven inputs to simulate incomplete real-world preferences, while weak supervision signals derived from these similarities guide the model’s learning.

A **LightGBM Ranker** trained under the **LambdaMART framework** optimizes ranking quality using pairwise relevance gradients, with performance evaluated via **NDCG (Normalized Discounted Cumulative Gain)** and interpretability assessed through feature importance analysis. Finally, during inference, the trained model ranks candidate properties for new users, generating consistent, interpretable, and preference-aligned housing recommendations.

### 6.3.1. Feature Mapping

Each mapping function measures how well a given property matches the user’s stated preference, regardless of whether the underlying feature is **numerical**, **categorical**, or **text-based**. For continuous features such as price or floor area, smooth Gaussian or saturating functions are used so that the similarity changes gradually rather than abruptly — allowing the ranker to learn continuous trade-offs between competing factors. For categorical or binary attributes (e.g., “new” vs “resale”), we defined consistent mapping rules that treat neutral or unknown preferences as moderate similarity rather than hard zero.

A special treatment was given to the **location feature**, which cannot be directly captured by a single scalar. Singapore is divided into 55 *Planning Areas*, and we computed a **spatial affinity matrix** based on the geographic distances between these regions. This spatial analysis allows the model to assess how acceptable a candidate property’s location is for a given user — properties closer to the user’s preferred area receive higher similarity scores. In this way, the system incorporates an important yet inherently qualitative factor, geographic proximity, into a quantitative scoring framework. Through these mapping functions, every user–property pair is converted into a vector of interpretable similarity features. This provides the foundation for later steps in the pipeline, where these similarity measures are used to generate weak labels and train the ranking model.

### 6.3.2. Training Sample Generation

The training data were built by systematically generating and labeling a set of user–property pairs. For each user, a group of candidate properties was first retrieved within a reasonable range of their budget, ensuring that the options were both financially relevant and sufficiently diverse. This retrieval process defined the base pool of potential matches from which the model would later learn individual ranking preferences.

Each retrieved property was then compared with the user’s profile through the feature mapping functions. These mappings produced a vector of continuous similarity values, denoted as **[sim\_\*]**, representing how well each property matched the user’s preferences across all key dimensions. To simulate real-world conditions—where users rarely specify every preference clearly—some user-driven features were randomly masked and replaced with neutral default values. Corresponding binary indicators, [**\*\_missing]**, were recorded to mark which similarities had been masked. This step ensured that the training process would reflect incomplete inputs and encourage the model to infer preferences even with partial information.

After computing the complete set of similarity features and missingness indicators, the system generated a continuous label to summarize the user’s overall affinity for each property. This weak label was formed through a weighted aggregation of the available similarity scores, where weights captured typical housing decision patterns: financial suitability, location, and living space were generally emphasized more heavily, while the remaining contextual factors followed the user’s declared priorities. The resulting value expressed how well a given property aligned with the user’s expected choices, even though it was not based on explicit ratings.

Finally, within each user’s candidate set, the continuous affinity scores were normalized and discretized into ordered relevance levels to facilitate learning-to-rank. Each final training record therefore consisted of **(user\_id, item\_id, sim\_ features, \_missing indicators, label)** that links user intent, property characteristics, and an approximate measure of relevance.

### 6.3.3. Model Training & Validation

The recommendation model was trained using **LightGBM Ranker**, which implements the **LambdaMART** algorithm — a gradient-boosted tree framework specially designed for **learning-to-rank** problems (1). **LambdaMART** extends traditional regression trees by optimizing a pairwise and listwise ranking objective that directly approximates ranking metrics such as the **Normalized Discounted Cumulative Gain (NDCG)**. This formulation allows the model to learn *relative preferences* rather than absolute scores, which aligns naturally with our housing-recommendation setting where we care primarily about the ordering of candidate properties for each user rather than predicting an absolute rating.

Each user’s candidate properties were treated as a single “query group” so the model learned to rank properties *within* each user’s candidate set rather than across users. The tree-based structure of **LightGBM** is particularly effective in our scenario: it handles heterogeneous inputs (continuous similarity features, binary missing-indicators), is robust to missing or neutral-filled features, and supports efficient training on tabular data with high interpretability through feature importance **(2).**

Formally, the **LambdaMART** loss can be expressed as:

where ρ denotes all item-pairs within the same user for which item is preferred to , is the model’s score prediction, and are gradient weights computed in proportion to the change in NDCG if the two items were swapped (3).

Training progress was evaluated using the **NDCG** metric on a validation set. NDCG is particularly suitable for learning-to-rank tasks because it measures both the relevance of items and their ordering positions, emphasizing the quality of the top-ranked results. This training design enabled the model to combine diverse similarity features and weak preference signals into a coherent ranking function that accurately captures user-specific housing preferences.

### 6.3.4. Inference and Application

During inference, the trained ranking model is used to generate personalized housing recommendations for new users. The inference pipeline mirrors the training process: candidate properties are first retrieved within a relevant budget range, similarity-based features are computed using the same transformation functions, and any missing user inputs are replaced with neutral defaults to ensure consistency with the training schema.

The model then assigns a relevance score to each candidate property, reflecting how well it aligns with the user’s stated preferences. These scores are used to rank all candidates, producing an ordered list of recommended flats. The final ranked output can be directly presented to end users or integrated into downstream housing search applications. This procedure ensures that recommendations remain consistent with the model’s ranking patterns and interpretable through feature-level similarity contributions.

**Database Layer**

The database layer serves as the foundational data infrastructure supporting all tools in the system. It integrates structured relational databases (SQL schemas) to enable efficient data storage, retrieval, and reasoning. Stored data encompasses user information, a comprehensive range of property-related attributes (e.g., geospatial location, historical pricing, transaction records), as well as multi-modal data such as property images and user-generated reviews collected from online sources. All data are either publicly available or gathered with explicit user consent, ensuring compliance with ethical and legal requirements.

# Implementation

**7.1 Technical Architecture and Stack**

The implementation adopts modular, multi-layered architecture designed for scalability, interpretability, and alignment with the project’s intelligent reasoning framework. Each layer performs a distinct role while maintaining seamless interoperability across the system.

7.1.1. LLM Layer (Core Agent)

At the language-model layer, the system’s cognitive core is powered by the GPT-5 API, with Google Gemini retained as a benchmarking and fallback option. The conversational agent is orchestrated through LangChain, chosen for its structured schema management, memory handling, and native tool-calling capabilities. Long-term contextual memory is managed using Mem0, which preserves user-specific histories and decision contexts across sessions. The agent maintains a Structured User Demand Schema in JSON format to capture constraints, preferences, and confidence levels, ensuring reproducibility and transparency. Crucially, the agent autonomously plans and executes the invocation of downstream analytical modules—each implemented as callable tools developed within this project—based on user prompts and reasoning objectives.

7.1.2. Tool Layer (Domain-Specific Modules)

This layer encapsulates the domain-specific analytical tools that operationalize the reasoning workflow. Each tool features clearly defined input–output contracts, enabling direct invocation by the LLM Agent.

* The HDB and the Condo Price Estimator leverages the XGBoost regression model trained on historical transaction data for predictive valuation.
* The Property Recommendation Engine, built using LightGBM Ranker under the LambdaMART framework, ranks candidate properties through preference-driven similarity features.
* Rule-based modules codify financial and policy checks (TDSR, MSR, LTV, BSD, ABSD, SSD) into version-controlled JSON/YAML rule sets.
* Spatial reasoning tasks employ OneMap API, GeoPandas, and Shapely for routing and accessibility computation.

7.1.3. Database Layer

The database layer forms the foundational backbone of the system, enabling persistent storage, structured retrieval, and high-performance access to both transactional and contextual data. In alignment with the design principles established in Section 6, this layer integrates multiple data sources—including HDB resale transactions, URA private residential records, and OneMap amenity datasets—into a coherent, queryable schema.

Structured data storage is managed through PostgreSQL, which organizes property attributes, transaction histories, and engineered features into relational tables with optimized indexing for temporal and spatial queries. Each record is associated with standardized identifiers such as postal codes and planning area codes, ensuring data integrity across analytical modules.

7.1.4. Orchestration and Infrastructure

The orchestration layer governs the coordination and execution of the system’s distributed components. It ensures that model pipelines, APIs, and reasoning workflows operate cohesively under unified control. Inter-component communication is managed through FastAPI, which exposes RESTful endpoints for model inference, rule evaluation, and recommendation services. The use of LangChain as the orchestration backbone enables the LLM Agent to manage tool invocation sequences, handle intermediate reasoning states, and ensure dependency resolution across the various analytical modules.

7.1.5. Interaction Layer

The LLM Agent mediates these interactions by interpreting the user’s intent, structuring the extracted information, and autonomously invoking the appropriate analytical tools such as the Price Estimator, Affordability Rule Checker, or Property Recommendation Engine.

Upon processing the query, the system presents outputs in a transparent and interpretable format. Users receive price estimates with confidence bands, affordability breakdowns (including TDSR and ABSD computations), and ranked property recommendations visualized through maps and charts. Each output is accompanied by contextual explanations, enabling users to understand how decisions were derived. This conversational front-end not only enhances accessibility but also reinforces the project’s goal of creating a transparent, explainable, and user-centric AI assistant for property decision-making.

# Expected Results and Progress

The project is expected to deliver an intelligent, demand-driven real-estate assistant powered by a large-language-model (LLM) agent. This agent will interpret user queries in natural language, extract requirements, and coordinate the relevant analytical modules to generate targeted outputs. It will manage both simple and complex multi-step tasks by invoking valuation models, recommendation engines, geospatial services, and rule-based calculators in combination.

For property sellers, the system will provide price predictions supported by estimated value ranges, confidence intervals, and explanations of key contributing factors. Comparable recent transactions will be presented as contextual evidence to enhance interpretability. For property buyers, the agent will generate personalized recommendations filtered and ranked by budget, location, accessibility, and lifestyle criteria, accompanied by affordability and regulatory checks that align with housing policies such as TDSR, LTV, BSD, ABSD, and SSD.

Users will also benefit from spatial visualizations, including route layouts, accessibility maps, and amenity insights that summarize commuting convenience, surrounding facilities, and neighborhood livability. Together, these outputs enable users to assess not only which properties meet their preferences but also how each option fits within broader financial and spatial contexts.

The LLM agent operates under a structured, rule-aware reasoning framework, ensuring that recommendations remain transparent, policy-compliant, and explainable. Each output is accompanied by justifications and references to maintain user trust. Through iterative dialogue, the system will clarify incomplete or ambiguous requests and deliver concise, goal-oriented responses.

Overall, the expected result is an intelligent, interactive, and user-centric reasoning system that unifies price estimation, policy validation, and recommendation within a single framework. It aims to reduce uncertainty, improve decision confidence, and create a transparent, data-driven experience for Singapore homebuyers and sellers.

# Challenges and Roadblocks

The development of the system is expected to encounter several technical and operational challenges. A primary limitation concerns data availability, as critical housing attributes such as unit orientation, renovation quality, or exact stack positioning are often unobservable in open datasets. This gap will be mitigated through the use of proxy features (e.g., floor level, building age, and facing direction) and the presentation of calibrated uncertainty bands in model outputs to maintain transparency. Another major consideration is policy drift, given that housing regulations such as ABSD, BSD, and TDSR are periodically revised. To ensure ongoing compliance, the system will incorporate a rule management layer that dynamically sources regulatory updates from official IRAS, MAS, and HDB repositories while maintaining version-controlled rule sets for auditability.

Generalization risk also poses a modeling challenge, as predictive accuracy may vary across regions or different phases of the property cycle. This will be addressed through stratified validation by town and market segment, as well as backtesting against historical downturns. Ethical safeguards are equally critical, since users may over-rely on automated valuations or recommendations for financially significant decisions. To mitigate this, outputs will include interpretability layers, uncertainty intervals, and explanatory disclaimers. Finally, qualitative or ambiguous user requests—such as “quiet neighborhood” or “near good schools”—will be handled through a structured User Demand Schema that maps soft preferences into quantifiable features for consistent system interpretation.

# Future Work

The current Minimum Viable Product (MVP) establishes the foundation of the system through three core capabilities: machine learning–based property valuation, amenity-aware recommendations, and personalized scenario analysis for buying and selling decisions. Building on this foundation, future development will extend the system’s functionality across adjacent domains of housing finance, protection, and lifestyle optimization.

* A Mortgage Refinancing Module will enable users to compare refinancing, repricing, and stay-put scenarios, projecting monthly instalments, cumulative interest, and break-even timelines after incorporating all associated fees.
* A Housing Insurance Advisor will guide users through policy selection for fire, renovation, and mortgage protection coverage, integrating a three-quote comparison engine for transparency and affordability.
* A Renovation Recommendation Suite will provide data-driven heuristics tailored to property type, floor area, and building age, suggesting value-enhancing scopes such as kitchen or bathroom upgrades, energy-efficient designs, and milestone-based budgeting. Over time, integration with partner networks and discount programs will further position the platform as a comprehensive, end-to-end decision assistant for homeowners.

# Conclusion

This project delivers an intelligent reasoning system designed to guide Singapore homebuyers and sellers through complex property upgrade and downgrade decisions. By integrating predictive valuation models, rule-aware affordability engines, and amenity-sensitive recommendation frameworks, the system moves beyond conventional search platforms to serve as a transparent, data-driven copilot for asset progression. It bridges the gap between market analytics and financial policy compliance, offering personalized insights grounded in both quantitative prediction and qualitative reasoning.

Looking ahead, the system has the potential to evolve into a comprehensive decision-making platform that continuously adapts to market dynamics and regulatory updates. Future extensions in mortgage refinancing, insurance comparison, and renovation planning will further strengthen its role as an integrated ecosystem for property lifecycle management. While challenges such as data granularity, policy drift, and ethical safeguards persist, the implementation of structured validation, version-controlled rule sets, and uncertainty-aware outputs ensures robustness and accountability. Ultimately, this project demonstrates how the convergence of machine learning, rule-based reasoning, and user-centric design can empower Singapore’s property owners to make confident, well-informed, and sustainable housing decisions.

# References

99.co. (2025). *Property Value Tool — Powered by SRX X-Value.*<https://www.99.co/singapore/property-value-tool>

Doshi-Velez, F., & Kim, B. (2017, February 28). Towards a rigorous science of interpretable machine learning [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1702.08608>

EdgeProp Singapore. (n.d.). *Edge Fair Value.*<https://www.edgeprop.sg/analytic/edgefairvalue>

EdgeProp Singapore. (n.d.). *“Is it a Good Deal?” (Deal Checker).*<https://www.edgeprop.sg/gooddealnew>

Gröbel, S., & Thomschke, L. (2018, July). Hedonic pricing and the spatial structure of housing data – An application to Berlin. Journal of Property Research, 35(3), 185–208. <https://doi.org/10.1080/09599916.2018.1510428>

Gunning, D., & Aha, D. W. (2019, June). *DARPA’s Explainable Artificial Intelligence (XAI) Program*. *AI Magazine, 40*(2), 44–58. [https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850](https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850?utm_source=chatgpt.com)

Housing & Development Board. (n.d.). *Application for an HDB Flat Eligibility (HFE) letter.*<https://www.hdb.gov.sg/residential/buying-a-flat/understanding-your-eligibility-and-housing-loan-options/application-for-an-hdb-flat-eligibility-hfe-letter>

Housing & Development Board. (2024, August 19). *Measures to cool the HDB resale market and provide more support for first-time home buyers.*<https://www.hdb.gov.sg/about-us/news-and-publications/press-releases/Measures-to-Cool-the-HDB-Resale-Market-and-Provide-More-Support-for-First-Time-Home-Buyers>

Housing & Development Board. (2025). *Resale price index from 1st quarter 1990 to 2nd quarter 2025 [Table].*<https://www.hdb.gov.sg/-/media/doc/EAPG-CSC/2Q2025-RPI-Table.pdf>

Housing & Development Board. (2025, April 1). *Upcoming flat supply and flash estimate of 1st quarter 2025 resale price index.*<https://www.hdb.gov.sg/about-us/news-and-publications/press-releases/Upcoming-Flat-Supply-and-Flash-Estimate-of-1st-Quarter-2025-Resale-Price-Index>

Housing & Development Board. (2025, April 11). *HFE letter application process.*<https://www.hdb.gov.sg/cs/infoweb/hdb-flat-portal/HFE/get-help/HFE-letter-application-process>

Housing & Development Board. (2025, July 3). *Credit to finance a flat purchase (incl. note on LTV 75%).*<https://www.hdb.gov.sg/residential/buying-a-flat/working-out-your-flat-budget/credit-to-finance-a-flat-purchase>

Housing & Development Board. (2025, July 23). *HDB launches 10,209 flats in the July 2025 BTO and SBF sales exercises [Press release].*<https://www.hdb.gov.sg/about-us/news-and-publications/press-releases/hdb-launches-10209-flats-in-the-july-2025-bto-and-sbf-sales-exercises>

Housing & Development Board. (2025, July 25). *Resale statistics (RPI Q2 2025 = 202.9).*<https://www.hdb.gov.sg/residential/selling-a-flat/overview/resale-statistics>

Housing & Development Board. (2025, July 31). *Outcome of HFE letter application.*<https://www.hdb.gov.sg/cs/infoweb/hdb-flat-portal/HFE/get-help/outcome-of-HFE-letter-application>

Inland Revenue Authority of Singapore. (2023). *Buyer’s Stamp Duty (BSD).*<https://www.iras.gov.sg/taxes/stamp-duty/for-property/buying-or-acquiring-property/buyer%27s-stamp-duty-%28bsd%29>

Inland Revenue Authority of Singapore. (2025, June 11). *Additional Buyer’s Stamp Duty (ABSD).*<https://www.iras.gov.sg/taxes/stamp-duty/for-property/buying-or-acquiring-property/additional-buyer%27s-stamp-duty-%28absd%29>

Inland Revenue Authority of Singapore. (2025). *Remission of ABSD for a married couple.*<https://www.iras.gov.sg/taxes/stamp-duty/for-property/appeals-refunds-reliefs-and-remissions/common-stamp-duty-remissions-and-reliefs-for-property/remission-of-absd-for-a-married-couple>

Inland Revenue Authority of Singapore. (2025, July 4). *Declaration form — SSD for residential properties [PDF].*<https://www.iras.gov.sg/media/docs/default-source/uploadedfiles/pdf/declaration-form-ssd-for-residential-properties.pdf>

Inland Revenue Authority of Singapore. (2025, July 4). *Seller’s Stamp Duty (SSD) for residential property.*<https://www.iras.gov.sg/taxes/stamp-duty/for-property/selling-or-disposing-property/seller%27s-stamp-duty-%28ssd%29-for-residential-property>

Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2024). Explainable Graph Neural Networks: *An Application to Open Statistics Knowledge Graphs for Estimating House Prices*. Technologies, 12(8), 128. <https://doi.org/10.3390/technologies12080128>

Ministry of National Development. (n.d.). *New flat classification framework: Standard, Plus, Prime.*<https://www.mnd.gov.sg/our-work/housing-a-nation/bto-classification>

Ministry of National Development. (2023, September 18). *Oral answer on the new classification framework for BTO flats.*<https://www.mnd.gov.sg/newsroom/speeches/view/oral-answer-by-ministry-of-national-development-on-the-new-classification-framework-for-bto-flats>

Monetary Authority of Singapore. (2021, December 15). *Measures to cool the property market [Media release & Annex].*<https://www.mas.gov.sg/news/media-releases/2021/measures-to-cool-the-property-market>

Monetary Authority of Singapore. (2021, December 15). *Mortgage servicing ratio and total debt servicing ratio rules [Explainer].*<https://www.mas.gov.sg/regulation/explainers/new-housing-loans/msr-and-tdsr-rules>

MOGUL.sg. (n.d.). *MAIA — AI property agent.*<https://www.mogul.sg/maia-homesearch>

Ohmyhome. (2025). *Homer Evaluation (HomerAI e-valuation).*<https://ohmyhome.com/en-sg/homer-evaluation>

PropertyGuru. (2025). *My Home — property valuation calculator.*<https://www.propertyguru.com.sg/my-home>

Sharma, H., Harsora, H., & Ogunleye, B. (2024, February 6). An Optimal House Price Prediction Algorithm: XGBoost. MDPI Analytics, 3(1), 30–45. <https://arxiv.org/abs/2402.04082>

Singapore Land Authority. (n.d.). *OneMap API documentation.*<https://www.onemap.gov.sg/docs/>

Singapore Land Authority (GeoWorks). (2025). *OneMap overview (APIs, POIs & routing).*<https://geoworks.sla.gov.sg/sla-products/onemap/>

SRX. (n.d.). *X-Value pricing calculator.*<https://www.srx.com.sg/xvalue-pricing>

Urban Redevelopment Authority. (2025, July 25). *Release of 2nd quarter 2025 real estate statistics [Media release].*<https://www.ura.gov.sg/Corporate/Media-Room/Media-Releases/pr25-40>

[1] Lyzhin I., Ustimenko A., Gulin A., Prokhorenkova L. (2022). *Which Tricks Are Important for Learning to Rank?* arXiv:2204.01500.   
[2] Ke G. et al. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* NeurIPS.   
[3] Fineis F. (2021). *The inner workings of the LambdaRank objective in LightGBM.* Blog.