# Machine Learning Canvas

PREDICTIONS

### End-user

Who will use the predictive system / who will be affected by it?

Potential Homebuyers, Property Sellers, Real Estate Agents and Brokers, Investors, HDB and Policy Makers, Marketing and Sales Teams

### Problem

Question to predict answers to (on behalf of user)

What will be the resale price of a specific HDB flat given its characteristics and market conditions?

Input (i.e. question "parameter")

Flat Characteristics, Location Data, Temporal Data

Possible outputs (i.e. "answers")

Predicted Resale Price

Type of problem (e.g. classification, regression, recommendation...)

Regression

Baseline: simple, alternative way of making predictions (e.g. manual rules)

Comparative Market Analysis (CMA)
Manual Rules

### Using predictions

When do we make predictions and how many?

- Prediction Timing:
- Predictions are made whenever a user inputs the characteristics of an HDB flat into the system. This could occur in real-time scenarios, such as during a property search by a potential buyer or a valuation request by a seller.
- Batch predictions might also be run periodically to provide market overviews or updates to stakeholders.
- Number of Predictions:
- The system could handle thousands of predictions daily, depending on user demand and batch processing needs.
- Real-time predictions are typically single or few per request, while batch predictions can cover all available listings or historical data points.

What is the time constraint for making those predictions?

Real-Time Predictions: These should be made within a few seconds to provide a seamless user experience, particularly when users are actively exploring or evaluating properties. Batch Predictions

OBJECTIVES

### Value proposition

What are we trying to do for the system's users? (e.g. spend less time on X, increase Y...)

Provide Accurate Price Estimates
Save Time and Effort
Enhance Market Understanding
Improve Market Efficiency

### Performance evaluation

Domain-specific / bottom-line metrics for monitoring performance in production

User Satisfaction
Transaction Volume
Revenue Impact

Prediction accuracy metrics (e.g. MSE if regression; % accuracy, #FP for classification)

Root Mean Squared Error (RMSE)
Mean Absolute Error (MAE)
$R^2$ (Coefficient of Determination)

Offline performance evaluation method (e.g. cross-validation or simple training/test split)

Cross-Validation
Training/Test Split

DATA

### Data sources

Where do/can we get data from? (internal database, 3rd party API, etc.)

Data.gov.sg
Google Maps API / OneMap API
HDB Official Websites and Publications
Real Estate Portals

### Data preparation

How do we get training data (inputs, and outputs if supervised learning)? How many data points?

Data Source: The training data is obtained from the comprehensive HDB resale transaction records available on data.gov.sg. This dataset includes historical resale prices and detailed flat characteristics from January 2017 to June 2024.

Data Points: The dataset contains approximately 181,000 rows, each representing a unique resale transaction.

Supervised Learning: This project is based on supervised learning, where the inputs are the features extracted from the dataset (such as flat type, floor area, location, and lease details), and the output is the resale price of the HDB flats.

Input features (extracted from data sources). If too many, list types of features and mention key ones.

Feature Types:
- Numerical Features: Quantitative values such as floor area in square meters, lease commencement date, remaining lease duration, and resale price.
- Categorical Features: Qualitative values that describe the flat and its location, such as town, flat type, flat model, storey range, block number, and street name.
- Temporal Features: Time-related data such as the transaction month.

Key Input Features:
- Flat Type: Classification of the flat based on the number of rooms (e.g., 3-room, 4-room).
- Floor Area (sqm): Total interior space of the flat in square meters.
- Flat Model: Design type of the flat (e.g., Improved, Model A).
- Town: The locality or district where the flat is located.
- Lease Commencement Date: The year the flat's lease started.
- Remaining Lease: Duration left on the lease, usually expressed in years and months.
- Storey Range: The range of floors where the flat is situated (e.g., 04 TO 06).
- Transaction Month: The month and year of the transaction, providing temporal context.

### Learning models

When do we create/update models? With which data / how much?

Model Creation and Update Frequency:
- Initial Model Creation: A baseline model is created using the entire historical dataset (from January 2017 to June 2024).
- Periodic Updates: The model should be updated periodically, for instance, monthly or quarterly, to incorporate new data and adapt to changes in the market.
- Event-Driven Updates: Models may also be updated in response to significant market changes or policy updates affecting HDB resale prices.

Data for Updates:
- Each update should use the most recent data available, ensuring that the model reflects the latest trends and conditions.
- Typically, all available data (including new and historical) should be used to maintain a comprehensive understanding of the market.

What is the time constraint for creating a model?

The initial training process may take several hours to days, depending on the data size and complexity of the model. Subse

How do we use predictions and confidence values?

Using Predictions:
- Predictions provide estimated resale prices for HDB flats, which can be used by buyers, sellers, and agents to guide decision-making.
- These estimates help in setting competitive prices, evaluating investment opportunities, and understanding market trends.

Using Confidence Values:
- Confidence values, or prediction intervals, give users a range within which the actual resale price is likely to fall.
- These values enhance user trust by providing transparency about the uncertainty or reliability of the predictions.
- They can be used to assess the risk associated with a given price estimate and to make more informed decisions.

Reset Form

Machine Learning Canvas v0.1



Criteria for deploying model (e.g. minimum performance value — absolute, relative to baseline or to previous model)

- Minimum Performance Threshold: The model should meet or exceed a predefined performance threshold, such as an RMSE less than SGD 30,000 or an $R^2$ greater than 0.80 on the test data.
- Improvement Over Baseline: The new model should perform better than the baseline or previous models, as measured by key metrics like RMSE, MAE, and $R^2$.
- Stability and Generalization: The model must demonstrate stability and the ability to generalize well across different subsets of the data, as verified through cross-validation or other evaluation methods.