**2025 - 2026 HOS Flat Price Prediction and Analysis System Report  
Author:** Alex Shang **Date:**26/2/2025

**Institutional Affiliation(s):** AIFT

## Table of contents

1. **Abstract**- [Page Number:3,4]
2. **Introduction** - [Page Number:5]
3. **Main Body**
   * Data Processing and Feature Engineering - [Page Number:6,7,8,9]
   * Model Construction and Evaluation - [Page Number:10,11,12]
   * Future Data Generation - [Page Number:13]
   * Interactive Visualization Interface - [Page Number:14,15]
4. **Conclusion** - [Page Number:16]

**Abstract:**

This report details a system designed for predicting and analysing HOS (Home Ownership Scheme) flat prices for 2025 - 2026. The system combines machine - learning techniques and an interactive visualization interface.

The data processing stage involves loading historical data from 2019 - 2024, followed by comprehensive feature engineering. Features such as time - series, market trends, lunar - related factors, and external interest rates are incorporated. The Random Forest Regressor is chosen as the prediction model, with parameter tuning via RandomizedSearchCV and TimeSeriesSplit for cross - validation. The best cross - validation R² score reaches 0.958, yet performance variability across folds hints at potential overfitting.

Future data for 2025 - 2026 is generated based on historical trends. The Streamlit - based interactive interface enables users to filter data by court/estate, price range, and floor range. It presents statistical information, comparison charts, and detailed price predictions, along with historical data visualizations like annual transaction volume and monthly price trends.

Despite its capabilities, the system has limitations. Overfitting risks persist, data quality could be enhanced, and future uncertainties are not fully accounted for. Recommendations for improvement include model optimization, further feature engineering, and real - time data updates.

**Introduction:**

In the dynamic landscape of the real estate market, accurate prediction of HOS flat prices is of great significance. For homebuyers, it provides crucial information for making informed purchasing decisions. Developers can leverage these predictions to plan future projects, and policymakers can use them to formulate relevant housing policies.

The 2025 - 2026 HOS Flat Price Prediction and Analysis System aims to address this need. By harnessing historical data and advanced machine - learning algorithms, it offers insights into future price trends. This report will comprehensively explore the system's components, from data handling and model building to the interactive interface and evaluate its performance and limitations.

**5. Main Body**

**5.1 Data Processing and Feature Engineering**

**5.1.1 Data Loading and Consolidation**

The system initiates by loading historical HOS sale - transaction data. Multiple JSON files, corresponding to different years (2019, 2021, 2022, 2023, with 2023 data also including 2024 information), are sourced from a specified directory. Each file contains records of transactions, which are then converted into DataFrames. These individual DataFrames are concatenated, resulting in a unified dataset. The data includes essential details such as the date of the agreement for sale and purchase, the name of the court/estate, floor number, saleable area of flats, and transaction price.

**5.1.2 Feature Engineering**

1. **Time - related Features**: To capture temporal patterns, the dataset is augmented with features derived from the transaction date. This includes extracting the year, month, and quarter. Additionally, the number of days since January 1, 2019, is calculated. Sinusoidal and cosine transformations of the month are also added, as they can potentially model seasonal patterns in the data.
2. **Market - trend Features**: Market - related features are engineered to reflect the changing price trends. For each court/estate, rolling mean values of the transaction price are calculated over different time windows. For example, a 6 - month rolling mean is used to represent the market trend, and a 3 - month lagged rolling mean is included to account for the delayed impact of market changes. The price per square meter is computed, and its rolling mean and standard deviation are also incorporated as features.
3. **Lunar Features**: Recognizing the potential influence of lunar - related factors on real estate transactions in some cultures, lunar features are added. The lunar month and an indicator for whether it is the lunar new year are included in the dataset. In case the lunardate library is not available, a fallback method is used to approximate these lunar features.
4. **External Data Incorporation**: External data, specifically real - interest - rate data, is integrated into the dataset. This data, which is provided on a monthly basis from 2019 - 2024, is merged with the main dataset based on the year and month. The interest rate is then forward - filled to ensure consistency.

**5.1.3 Data Validation and Cleaning**

The dataset undergoes a rigorous validation and cleaning process. Critical columns such as the transaction price and saleable area of flats are checked for missing values, and rows with missing data in these columns are removed. An outlier detection method based on the median absolute deviation (MAD) is applied to the price - per - square - meter data. Additionally, the saleable area of flats is filtered to ensure it lies within a reasonable range (between 5 and 500 square meters). The floor number is converted to a numeric type, with missing values filled with 0.

**5.2 Model Construction and Evaluation**

**5.2.1 Model Selection**

The Random Forest Regressor is selected as the core prediction model. This algorithm is chosen for its ability to handle complex, non - linear relationships in the data. It aggregates multiple decision trees, which helps in reducing overfitting and improving generalization performance. The model is initialized with a random state of 42 for reproducibility and set to use all available CPU cores (n\_jobs=-1) for faster processing.

**5.2.2 Model Training and Tuning**

A parameter grid is defined for the Random Forest Regressor, which includes parameters such as the number of estimators (n\_estimators), maximum depth (max\_depth), minimum samples per leaf (min\_samples\_leaf), minimum samples for split (min\_samples\_split), and maximum features (max\_features). RandomizedSearchCV is employed to search for the optimal combination of these parameters. This method randomly samples parameter combinations from the defined grid and evaluates the model using TimeSeriesSplit cross - validation. TimeSeriesSplit is used to ensure that the model is tested on data that follows the temporal order, which is crucial for time - series - related data.

**5.2.3 Model Evaluation**

The model's performance is evaluated using several metrics. The best cross - validation R² score obtained is 0.958, indicating that the model can explain a significant portion of the variance in the data. However, when looking at the performance across different folds of the TimeSeriesSplit, there is variability. For example, Fold 2 has an R² score of 0.845, while Fold 5 has an R² score of 0.983. This variability suggests that the model may be overfitting to some specific subsets of the data. Other metrics such as root - mean - squared error (RMSE) and mean - absolute error (MAE) also show differences across folds, further highlighting the potential overfitting issue.

**5.3 Future Data Generation**

To generate future data for 2025 - 2026, the system follows a multi - step process. First, a template DataFrame is created for each unique court/estate. This template is based on the latest available data for each court/estate in the historical dataset. Then, future dates are generated from January 1, 2025, to December 31, 2026, at a monthly frequency.

For each future date, time - related features are calculated in a similar way to the historical data. Market - related features are predicted by assuming a certain growth rate. The annual growth rate of the transaction price is calculated from the historical data, and this rate is used to project future values of market - trend - related features such as the market trend, price per square meter, and their lagged versions.

External factors, such as the interest rate, are also projected. The average interest rate in 2024 is used as a base, and a simple growth factor is applied to predict future interest rates. All these generated features are combined to form the future dataset, which is then used for price prediction.

**5.4 Interactive Visualization Interface**

**5.4.1 Interface Design**

The interactive visualization interface is developed using the Streamlit framework. The interface has a clean and intuitive design. At the top, a title is displayed, which clearly states the purpose of the system. A spinner is shown during the data - loading process to inform the user.

The sidebar of the interface serves as a filter panel. Users can select specific court/estates, price ranges, and floor ranges. A "Reset Filters" button is provided to quickly reset all filters. Based on the user's selections, the main content area of the interface displays relevant information.

**5.4.2 Visualization Charts**

1. **Annual Transaction Volume Chart**: This bar chart shows the number of transactions in each year. It helps users understand the overall activity level of the HOS market over the years. The x - axis represents the years, and the y - axis represents the transaction volume.
2. **Monthly Price Trend Chart**: A line chart that plots the average monthly price. It includes both historical and predicted prices. The historical prices are shown in one color, and the predicted prices for 2025 - 2026 are shown in another color. This chart allows users to observe the long - term price trends and how the predicted prices fit into the historical pattern.
3. **Court/Estate Comparison Chart**: When users select multiple court/estates, this bar chart compares the average predicted prices of the selected court/estates. It helps users quickly identify the differences in price levels among different properties.

**6. Conclusion**

The 2025 - 2026 HOS Flat Price Prediction and Analysis System represents a significant effort in leveraging historical data and machine - learning techniques to predict future HOS flat prices. Through comprehensive data processing, feature engineering, and model construction, the system has achieved a relatively high cross - validation R² score, indicating its ability to capture patterns in the historical data.

The interactive visualization interface further enhances the usability of the system, allowing users to explore different aspects of the data and predictions. However, the system is not without its limitations. The variability in model performance across different folds of cross - validation points to potential overfitting issues. Additionally, the data quality could be further improved, and the model may not fully account for the complex and unpredictable factors that can influence future real estate prices.

Despite these limitations, the system provides valuable insights into the potential price trends of HOS flats in 2025 - 2026. It can serve as a useful tool for homebuyers, developers, and policymakers in making more informed decisions.