Model Prediction Task 1

OBJECTIVE

- To build a predictive algorithm to determine the factors affecting prices of residential properties in Singapore
- To identify potential strategies in curbing housing prices inflation

PROCESS

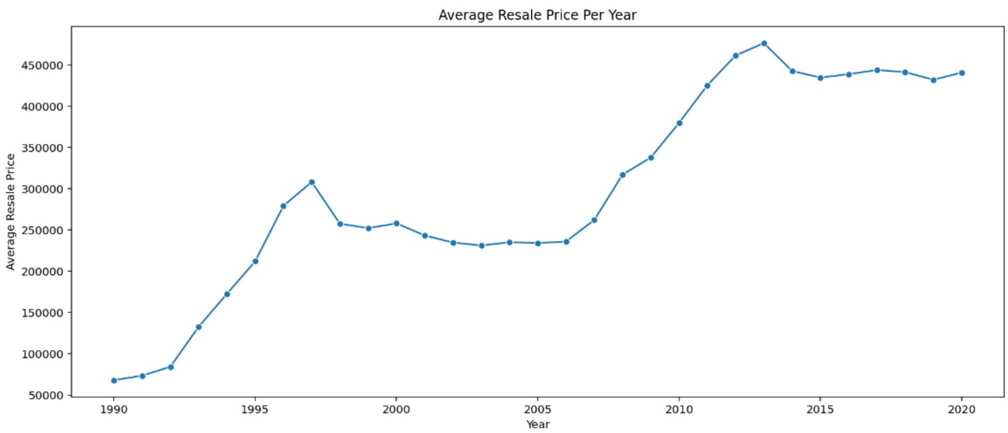
1. Data Preparation
2. Data Exploration
3. Data Pre-processing
4. Model Selection and Training
5. Tuning and Validation
6. Iteration

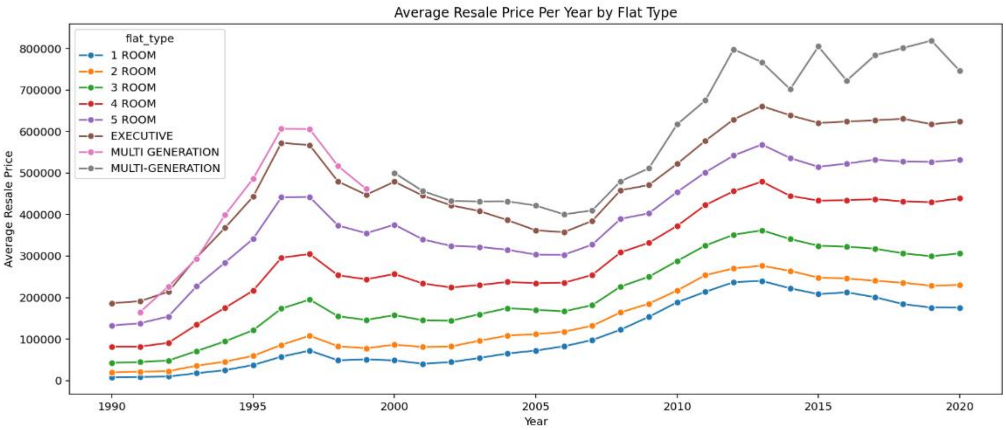
Data Preparation

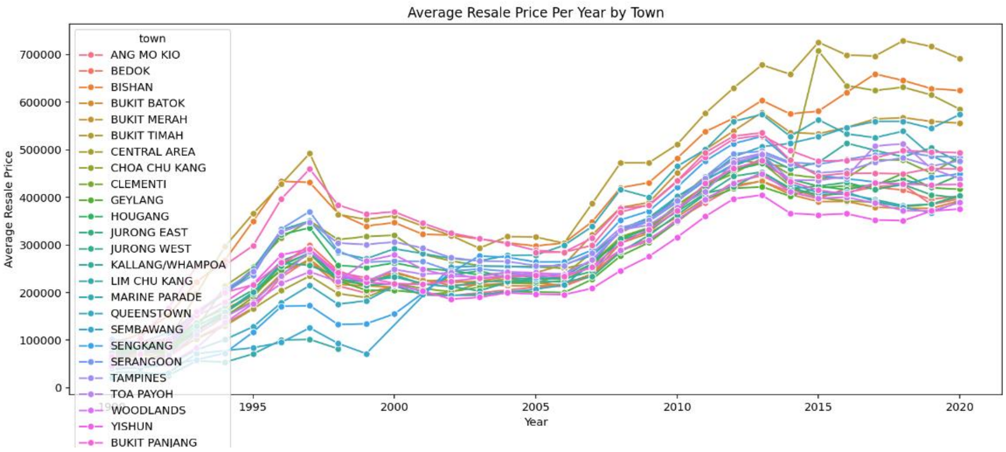
- Consolidation and organisation of dataset
- Cleaning of data types
- Check for missing or null values

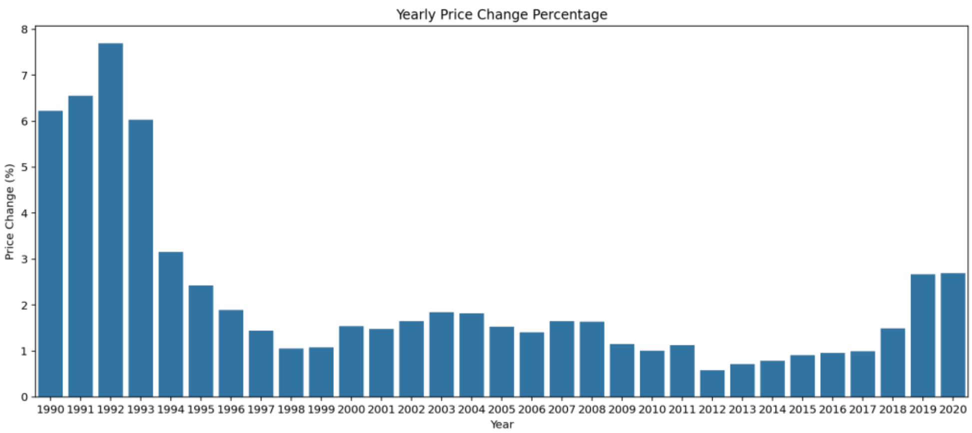
Data Exploration

- Exploratory data analysis to investigate main characteristics of dataset









Data Pre-processing

- Smaller subset of data used for initial experimentation and tuning
    - Most recent data was used—after significant property price cooling measures in 2018[1]

[1] https://www.channelnewsasia.com/singapore/property-cooling-measures-hdb-resale-prices-2013-2018-each-singapore-town-2385831

- Number of features reduced to reduce computational requirements
    - Aim to identify a good set of parameters, before training the final model on the full dataset
    - However, with access to more computational resources, parallelised search across multiple machines could be done
- Imputation and one-hot encoding for categorical variables

Model Selection

- Random Forest Regressor was chosen for this task due to its following advantages:
    - Ability to handle both numeric and categorical features
    - Robust to outliers and non-linear relationships
    - Provides feature importance scores
    - Generally performs well on a variety of datasets without extensive tuning
- Initial tests with (n_estimators = 300, max_depth = 5, random_state = 42), yielded the following metrics:
    - Test Set Root Mean Square Error (RMSE): 88934.04
    - Test Set R-squared (R^2): 0.66

Hyperparameter Tuning

- BayesSearchCV was used to perform an optimised cross-validated search over a predefined parameter grid. A fixed number of parameter settings is sampled from the specified distributions. Search parameters were as follows:
    - Number of trees (n_estimators): (100, 1000)
    - Maximum tree depth (max_depth): (10, 100)
    - Minimum number of samples required to split an internal node (min_samples_split) (2, 10)
    - Minimum number of samples required to be at a leaf node (min_samples_leaf): (1, 5)
- BayesSearchCV chosen over the more comprehensive GridSearchCV (in which all parameter values are tried out), given the limitations in computational resources – less time, fewer iterations required
- BayesSearchCV also more efficient than RandomizedSearchCV
- BayesSearchCV yielded the following results:
    - Best parameters: [('max_depth', 31), ('min_samples_leaf', 1), ('min_samples_split', 10), ('n_estimators', 1000)]
    - Best score: 0.9338171869649884

RESULTS

Model Performance Metrics

- Root Mean Square Error (RMSE): 40647.39
- R-squared (R^2): 0.93

These metrics indicate that the model explains 93% of the variance in housing prices and has an average prediction error of SGD$40,647.39. This is a significant improvement over the initial test scores.

Feature Importance

Top 5 most important features and their Gini importances are:

1. floor_area_sqm (0.412663)

2. remaining_lease (0.128817)

3. flat_type_4 ROOM (0.091736)

4. town_BUKIT MERAH (0.051083)

5. re_binned_range (0.049248)

INSIGHTS

Key Factors Affecting Housing Prices

1. Floor Area: Larger units generally command higher prices, reflecting the premium placed on space in Singapore's urban environment.

2. Remaining Lease: Properties with longer remaining leases tend to have higher values, indicating buyers' preference for newer or recently renewed leases.

3. Location: Certain towns (Bukit Merah, Queenstown, Bishan, Central Area, Toa Payoh) consistently show higher property values, likely due to factors such as proximity to the city centre, amenities, and transportation links.

4. Flat Type: The type of flat does influence prices, with 4-room flats generally commanding higher prices.

5. Floor Level: Higher floor levels often correlate with higher prices, possibly due to better views and ventilation.

Potential Strategies for Curbing Housing Price Inflation

1. Targeted Development: Focus on developing more housing in areas with lower median prices to increase supply in these regions and potentially alleviate pressure on high-demand areas.

2. Lease Renewal Programs: Implement programs to facilitate lease renewals or extensions, which could help stabilise prices for older properties.

3. Size-Based Pricing Policies: Introduce policies that encourage a balance between unit sizes, potentially by adjusting pricing or grant structures based on floor area.

4. Amenities and Infrastructure Development: Improve connectivity to areas with lower housing prices and develop local amenities, potentially increasing their attractiveness and distributing demand more evenly.

5. Adaptive Pricing for New Flats: Use the insights from this model to inform pricing strategies for new Build-To-Order (BTO) flats, ensuring they remain competitive and accessible.

LIMITATIONS

Model Limitations

- Model does not account for macroeconomic factors such as interest rates, GDP growth, or employment rates, which can influence housing prices
- Bayesian Search is not as comprehensive as Grid Search
- Temporal factors (e.g., seasonal trends, long-term market cycles) are not explicitly modelled
- Impact of nearby amenities (schools, shopping centres, parks) is not directly captured in the available features

Suggestions for Improvement

1. Train and test model with larger dataset
2. Incorporate additional data sources:
    a. Macroeconomic indicators
    b. Proximity to amenities (schools, MRT stations, shopping centres)
    c. Urban development plans
3. Explore time series modelling to capture temporal trends and seasonality
4. Investigate the use of geospatial models to better capture location-based effects
5. Dedicate sufficient time and resources to model development

REPOSITORY

https://github.com/dyhq/sg-housing-price-prediction