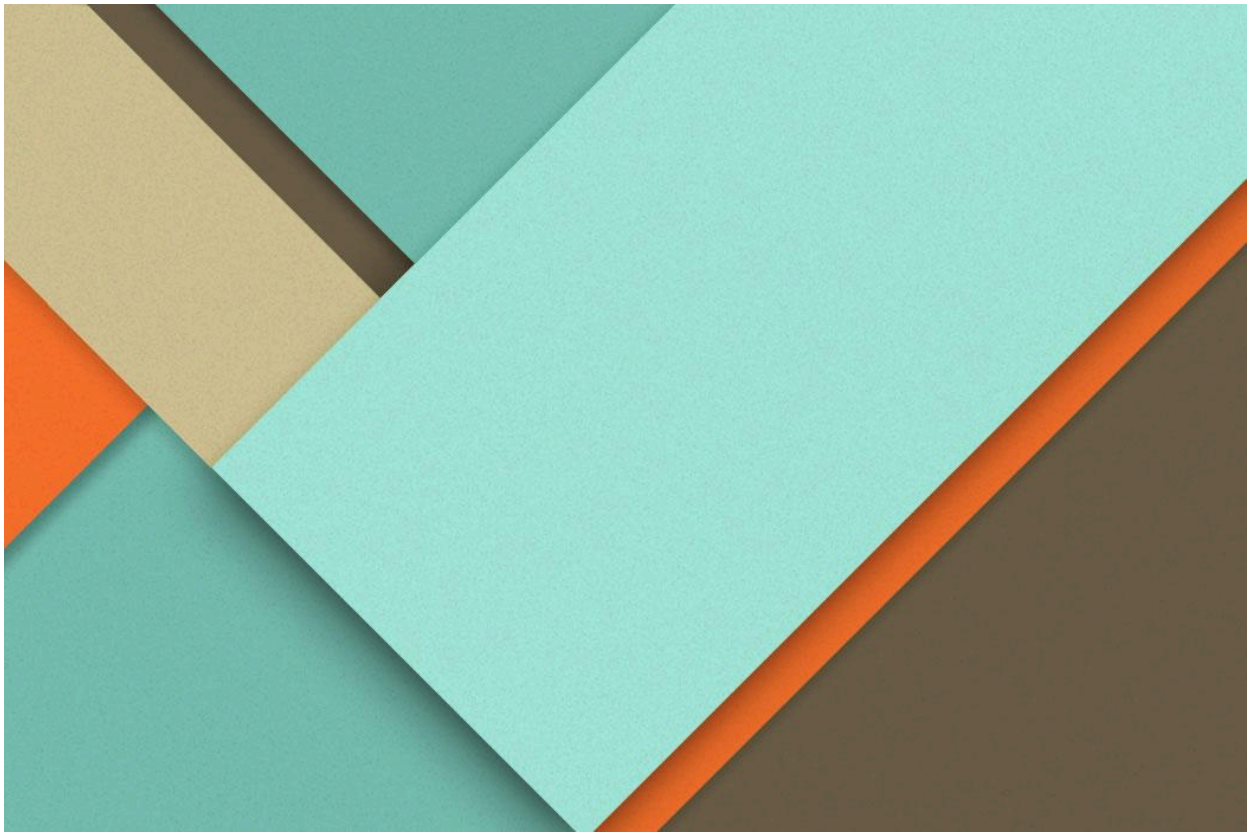


Predicting Housing Resale Prices

08.04.2025

—

By Home Bros:
Quoid Jing-Xuan Andrel, Jinal Surana, Shah Aashka Anand, Asha Meredith Solomon, Saamiya Khan

Overview

We will be utilizing datasets from data.gov.sg containing true values of housing resale prices, as well as other continuous attributes to conduct supervised learning to predict housing resale prices.

Data Collection

- Resale Flat Prices | HDB | data.gov.sg
- Real GDP
- Unemployment Rate (End Of Period), Quarterly, Seasonally Adjusted | SINGSTAT | data.gov.sg
- Population Size

Data Preprocessing

- The attributes we will be referencing from the datasets are:
    a. Area Square Meters
    b. Flat Type
    c. Month of Sale
    d. Remaining Lease
    e. Real GDP (annual)
    f. Flat Model
    g. Population
    h. Unemployment rate
    i. Distance to CBD

Feature Engineering

- We will use the scikit-learn Standard Scaler

Model Selection & Training

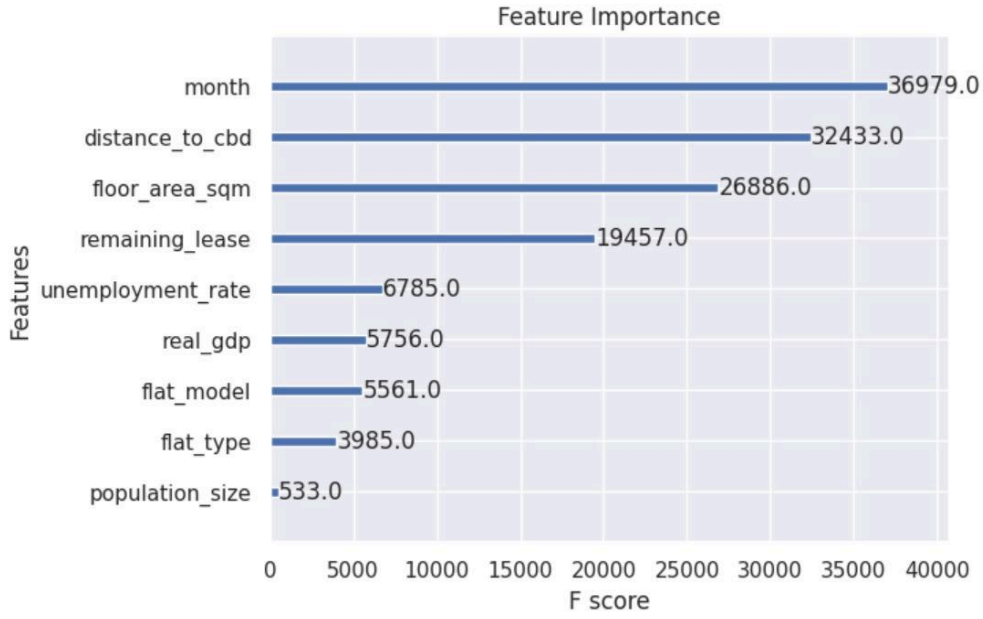
- Linear Regression
- RandomForestRegressor
- XGboost Regressor
- Optuna/gridsearchcv (hyperparameter tuning)

Model Evaluation

- Test/Validation/Training split of 70/15/15 to be used
  (Data from 1990-2009 to be trained on, 2010-2014 for hyperparameter tuning and 2015-2019 to be tested on)

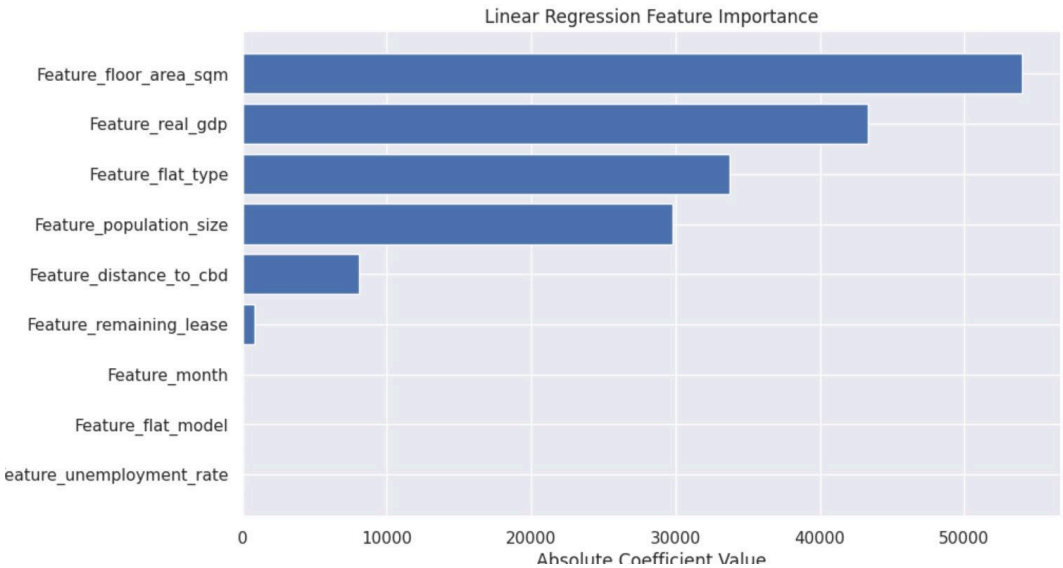
- Linear Regression had the highest MAPE among the 3 models and thus it was the least preferred
- Results from XGboost Regressor: MAPE was 11% after retraining the model on combined validation and training sets. The top 2 important features for this model were the month and the distance of the flat to CBD.

```
Fitting 3 folds for each of 27 candidates, totalling 81 fits
Best parameters: {'learning_rate': 0.2, 'max_depth': 10, 'n_estimators': 100}
Best score (Neg MAE): -46531.43847289029
Validation MAPE 0.18852035284363808
Validation MAE: 85271.21010429255
Validation RMSE: 102414.62
Test MAPE 0.18861463931502254
Test MAE: 92935.54742497583
Test RMSE: 122875.79
Test MAPE after retraining 0.11016206176608687
Test MAE after retraining: 46290.441612714356
Test RMSE after retraining: 63781.09
```



- Results from Linear Regression: MAPE was 25% after retraining which was an increase from the initial 15% suggesting model overfitting during hyperparameter tuning. The top 2 important features for this model were the flat's floor area in square meters and the real gdp in the period the time the flat was sold.

```
Fitting 3 folds for each of 16 candidates, totalling 48 fits
Best parameters: {'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'positive': True}
Best score (Neg MAE): -72599.63462787688
Validation RMSE: 115727.92
Validation MAE: 89299.10
Validation MAPE: 0.1954
Test MAPE 0.1585726539284444
Test MAE: 81838.68671128732
Test RMSE: 128268.20
Test RMSE after retraining: 122703.64
Test MAE after retraining: 100550.04
Test MAPE after retraining: 0.2497
```



- Results from RandomForestRegressor:

Desired Outcome

- Lowest possible MAE (mean absolute error)
- Analysis of MAE across years/towns
- Feature Importance Ranking (determine which feature affects the prediction most) - SHAP Analysis

Work Distribution

I. Team 1

Saamiya, Aashka, Andrel : To work on Linear Regression and XGboost models.

II. Team 2

Asha, Jinal : To work on the RandomForest Regression model.

Team 1 timeline

I. April 20th

Data preprocessing (Asha, Saamiya)

II. April 23rd

Linear regression

III. April 27th

XGBoost

IV. April 30th

Model Evaluation

V. May 6th

Project file update (Aashka)

Final Submission

Team 2 timeline

I. April 26th

Functioning model to be completed, training to start

II. April 30th

General/Large adjustments should be completed, fine tuning to begin

III. May 6th

Final Submission