Report

UNIVERSITY OF WOLLONGONG AUSTRALIA

Singapore Institute of Management

Predicting Resale HDB Prices With Machine Learning Models

CSCI323 - Modern Artificial Intelligence

By Group 11

Members:

Student's Name	UOW ID
Lester Liam Chong Bin	7558752
Bryce Nicolas Fernandez Sumcad	8561369
Jesyln Ho Ka Yan	8535383
PARK KI SUNG	8379129
Lee Donghyun	8876320
Chea Darayuth	8550864

Executive Summary

This report presents a machine learning approach to forecasting the resale prices of Housing Development Board (HDB) flats in Singapore using a structured dataset of over 200,000 transactions from January 2017 to May 2025. By analyzing attributes such as town, flat type, floor area, and lease details, we developed and evaluated four models: XGBoost, Random Forest, Multilayer Perceptron (MLP), and Skforecast. The objective was not only to predict prices but also to assess how various features influence market trends. XGBoost emerged as the most reliable model, consistently outperforming others in both RMSE and R^2 metrics, especially after applying feature engineering. Our findings also highlight critical challenges such as data imbalance across towns and the absence of geospatial and exogenous variables. These insights inform future recommendations on improving model accuracy, including integrating location-based features and exploring ensemble strategies to strengthen predictions in low-volume regions. This work demonstrates the potential of AI in real estate analytics and provides a foundation for more context-aware, data-driven decision-making in Singapore's housing market.

Table of Contents

1.0 Introduction

This project focuses on applying artificial intelligence, specifically machine learning, to predict the resale prices of HDB flats in Singapore. Using a dataset of over 200,000 resale transactions recorded between January 2017 and May 2025, we explore how various factors—such as flat type, location, floor area, and lease details—affect pricing patterns. The dataset includes 11 key attributes, including the transaction month, town, floor area, flat type, lease commencement year, remaining lease, and final resale price. With this rich information, we apply machine learning techniques to uncover patterns and relationships between different housing features and market value.

Our main objective is to build predictive models that can estimate resale prices accurately and analyse the most influential factors driving those prices. Since property values are shaped by a combination of structural, geographical, and temporal elements, this project sits at the intersection of real estate analytics, urban development, and AI-driven forecasting. We believe this work has wide-reaching applications. In the property market, it can help buyers and sellers assess fair prices and spot potential deals. Real estate platforms might integrate these models into their pricing tools or dashboards to enhance user experience. Policymakers can also use the insights to better understand housing trends and develop data-informed strategies to maintain affordability. Lastly, the project offers a valuable case study in applying machine learning to real-world structured data, making it a practical learning resource for students and educators in AI.

2.0 Background Theory

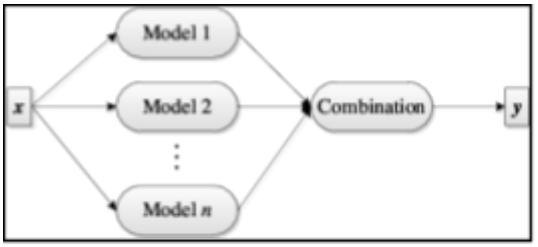
2.1 Theoretical Foundation of the Topic

House price prediction is a well-established application of supervised machine learning, traditionally modeled as a regression task. The objective is to estimate property prices based on features such as flat type, floor area, location, storey range, and lease remaining. These features introduce both linear and nonlinear interactions, which calls for models capable of capturing complex patterns beyond simple regression.

2.2 Theoretical Development in Existing Research

XGBoost & Random Forest:

Previous studies conducted by Wang et al. (2016), have shown that ensemble-based methods like XGBoost and Random Forest outperform linear models in predicting HDB resale prices due to their ability to capture nonlinear feature interactions and handle heterogeneous data types.



(Michael Affenzeller 2015)

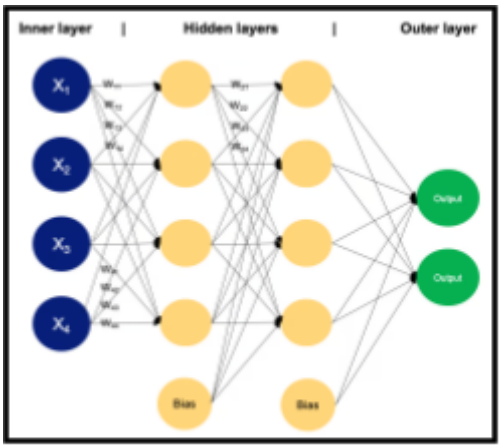
- **Random Forest** aggregates multiple decision trees trained on bootstrapped samples using the bagging technique. Its final output is the average of predictions from all trees.
    - **Architecture:** Input features → Multiple Decision Trees → Averaged Output
    - **Key formula:**

      Where $h_t(x)$ is the prediction of the t^{th} tree. $$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$
- **XGBoost** improves upon gradient boosting by incorporating regularization and efficient tree pruning to prevent overfitting. It builds trees sequentially, each correcting the errors of its predecessor.

  Architecture: Input → Additive Trees with Gradient Boosting → Final Prediction
    - **Key formula:** Where F is the space of regression trees and f_k are the functions learned. $$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

Multilayer Perceptron (MLP):

Neural networks have also been explored for housing price prediction Jaiswal (2025). MLPs are feedforward networks consisting of one or more hidden layers with non-linear activation functions, enabling them to approximate any continuous function.



(Jaiswal 2025)

Architecture:

- Input → Hidden Layers (ReLU/Tanh) → Output Layer (Linear)

Mathematical Representation:

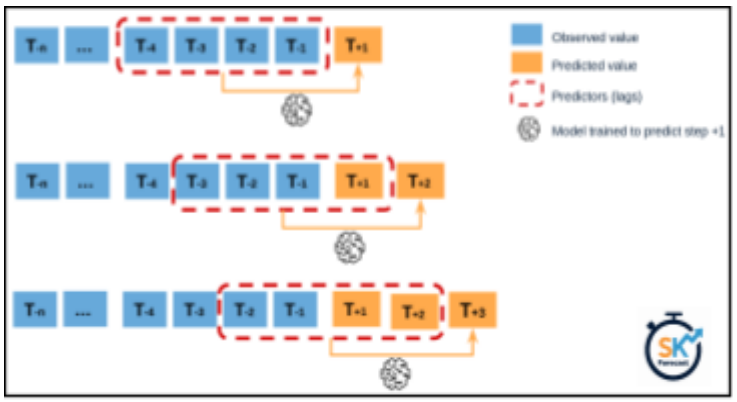
- Single Hidden Layer: where ϕ is a non-linear activation function (e.g., ReLU), and W, b are weights and biases.

$$\hat{y} = W^{(2)} \cdot \phi(W^{(1)}x + b^{(1)}) + b^{(2)}$$

Skforecast & Time Series Modeling:

While resale flat prices are not purely time-dependent, some trends do exhibit temporal patterns. According to (Skforecast, 2024), Skforecast is a wrapper that allows autoregressive modeling using regressors like XGBoost or Random Forest. It incorporates lagged variables as predictors to learn from historical trends.

- **Application in prior work:** Used to forecast sequential values like sales volume or average monthly prices, capturing seasonality or trend shifts.
- **Architecture:**
  Time series y_t → Feature engineering with lag values → Regression model → Prediction
- **Mathematical Formulation:** $\hat{y}_{t+h}$
  Where f is a supervised regression mode $\hat{y}_{t+h} = f(y_t, y_{t-1}, \cdots, y_{t-n})$



(Skforecast, 2024)

2.3 Fit with Our Group's Design and Solution

Our goal for this project is to accurately train, and predict Resale HDB Flat Prices using various machine learning models. Additionally, we want to be able to predict future flat prices.

Based on our research, we will use the following models and techniques:

- **Random Forest Regressor**: Great baseline model for complex features and patterns
- **XGBoost Regressor**: Superior performance and predictive capabilities with gradient boosting
  MLP Regressor: Explores deep learning for capabilities complex non-linearities.
- **Skforecast**: Alternative approach, using time series based modelling

Our evaluation strategy will be using primarily Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). According to Matalonga (2023), she states that datasets with a high number of outliers should use MAE instead of RMSE as it's less sensitive to outliers. Hence, given our prediction is focused on future values, these metrics are appropriate for regression problems as they provide a measure of the model performance, for example a more sensitive model may be beneficial for riskier forecasting, while MAE may be more suitable if we want a more interpretable model instead.

3.0 Solutions, Evaluation, and Discussion

3.1 Summary of Dataset Observations

Exploratory Data Analysis (EDA) revealed that resale flat prices generally increased over time with some seasonal variation. Larger flat types (e.g., Executive, 5-room) had higher median prices. Towns like Queenstown and Bishan showed significantly higher average prices, likely due to their central locations and estate maturity. Additionally, a clear inverse relationship between remaining lease and price was observed, validating its role as a key predictive feature.

3.2 Design and Implementation

Training Setup: The dataset was split based on transaction dates to mimic a real-world forecasting scenarios:

- **Training:** **2017-01 → 2024-05**
- **Validation:** **2024-06 → 2024-12**
- **Testing:** **2025-01 → 2025-05**

This forward-looking split simulates how models perform when trained on past data and used to predict future unseen records, which is especially relevant in time-sensitive housing markets like HDB Resale.

3.2.1 Data Transformation:

skforecast: To enable time series forecasting, the dataset was aggregated to a monthly time series basis, and filled with average resale price for by year, month, town, flat type, if a value does not exist we reduce to by year, town, flat type, lastly by year and flat type. Lastly, we transposed the transposed into multiple columns series, where each column represent a time series for 1 HDB Town Flat Type *(eg: "ANG MO KIO 3 ROOM")*

3.2.2 Preprocessing:

We cleaned the dataset by removing missing values and duplicates to ensure data integrity. Due to the high number of unique categories present in the dataset. Categorical features such as *town, flat_type, flat_model* were encoded using BinaryEncoder, numerical features like *floor_area_sqm* and *remaining_lease* were scaled using RobustScaler. However, time related data like *month* and *year* were allowed to remain.

3.2.3 Feature Selection:

We've decided to include all features, except for *block (2741 labels)* and *street_name (574 labels)* which were dropped from the dataset. This decision was made based on two main factors:

1. **Extremely High Cardinality & Inconsistency**: There are many unique combinations of street names, which can lead to sparse data and overfitting.
2. **Duplicate Identifiers**: The same block number can appear in multiple towns (e.g., Block 123 exists in both Tampines and Bedok), making them unreliable as distinct predictive features.

This selection process helps to reduce dimensionality, minimize overfitting risk, and enhance model generalizability.

3.2.4 Feature Engineering:

To improve model performance, we've integrated external data sources such as Annual Inflation Rate (%) for Consumer Prices by World Bank Data, HDB Resale Price Index, and HDB Demand for Rental and Sold Flats to capture macroeconomic influences. We also engineered lag features such as *prev_year_mean_price* on past transactions to help the model learn time-dependent patterns. These steps enhanced the Random Forest model's ability to understand both temporal and contextual factors in resale price prediction.

3.3 Observations from Exploratory Data Analysis (EDA)

Our EDA revealed clear patterns, imbalance dataset in HDB resale activity:

- Towns like **Sengkang, Punggol, Woodlands** had the highest transaction volumes, likely due to their mature infrastructure and large HDB populations. In contrast, areas such as **Bukit Timah** and the **Central Area** showed fewer sales due to limited public housing supply.
- **Units Sold per Flat Type**: **4-room flats** were the most commonly resold across all towns, reflecting strong demand from families. **3-room flats** followed, while **1-room, 2-room, Executive and Multi-Generation** flats saw fewer transactions, possibly due to higher prices or limited availability.

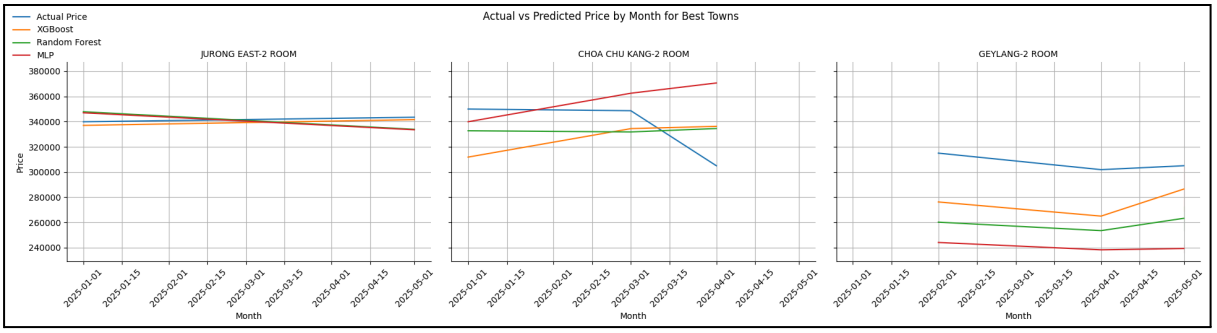
3.4 Evaluation Results and Discussion

The performance metrics across four models are summarized below:

Model	RMSE		MAE		R^2	
Feature Engineering	✗	✔	✗	✔	✗	✔
Random Forest	75951.33	76095.03	57018.58	55121.53	0.8554	0.8546
XGRegresssor	70078.48	57286.15	54476.83	42405.33	0.8769	0.9176
MLP Regressor	123756.33	71531.05	84953.94	50618.24	0.6161	0.8715

Among all models, **XGBoost achieved the best performance**, achieving the lowest RMSE (57286.15) and MAE (42405.33) and the highest R^2 (0.9176), indicating superior accuracy and a strong fit to the data.

Other Observations: Our evaluation of predicted and actual prices across towns revealed the following insights for each town flat type:



The XGboost model predictions had the closest predicted with the actual prices as shown in the diagram above across the "Best Towns" shown, especially for Jurong East-2 Room and Geylang-2 Room. The Random Forest model underestimated significant prices for Geylang-2 Room, and the MLP model showed significant both overestimate and underestimate.

3.5 Model tracking and Feature Engineering Evaluation

Our evaluation across models offered the following observations:

Prediction Trends Over Time	Impact of Feature Engineering
Over time, the XGBoost model showed the most consistent predictions with actual resale price trends across months in towns like Choa Chu Kang, Jurong West, and Bukit Merah. Its predictions closely tracked actual market trends, but Random Forest and, particularly, MLP showed higher variances over time.	Implementing Feature Engineering resulted in clear performance improvement. Compared to the application without feature augmentation, all models, particularly XGBoost and Random Forest, showed lower prediction error and closely followed the actual price patterns across months.

Overall Performance Results:

The performance of each model was evaluated using **R^2**, **RMSE**, and **MAE** on the test dataset, and additional RMSE values were calculated ***across different towns*** and ***flat types*** to assess model accuracy at a more localized level.

Without Feature Engineering

XGBoost achieved:
R^2: 0.8769**, MAE:** 54,476.83, **RMSE:** 70,078.48
Random Forest achieved:
R^2: 0.8554**, MAE:** 57,018.58, **RMSE:** 75,951.33
MLP achieved:
R^2: 0.6161, **MAE:** 84,953.94**, RMSE:** 123,756.33

town	flat_type	metric_name	value
JURONG EAST	2 ROOM	xgb_rmse	16917.388051
CHOA CHU KANG	2 ROOM	rf_rmse	18439.938664
GEYLANG	2 ROOM	mlp_rmse	22812.274236
TAMPINES	2 ROOM	xgb_rmse	23333.176973
SEMBAWANG	2 ROOM	xgb_rmse	24103.567370

These metrics clearly indicate that **XGBoost** and **Random Forest** outperform **MLP**, especially in high-density towns and common flat type.

With Feature Engineering

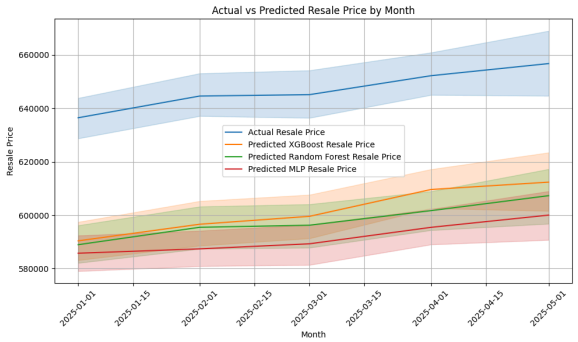
XGBoost
R^2: 0.9176**, MAE:** 42405.33**, RMSE:** 57286.15
RandomForest achieved
R^2: 0.8546, **MAE:** 55121.53, **RMSE:** 75095.03
MLP achieved
R^2: 0.8683**, MAE:** 51694.62, **RMSE:** 72411.03

town	flat_type	metric_name	value
JURONG EAST	2 ROOM	xgb_rmse	5777.237266
TAMPINES	MULTI-GENERATION	mlp_rmse	6401.250000
JURONG EAST	2 ROOM	rf_rmse	10528.394899
JURONG EAST	2 ROOM	mlp_rmse	10533.016942
YISHUN	MULTI-GENERATION	rf_rmse	13811.365000

The performance metrics highlight the best-case **RMSEs** for each model across specific **town–flat type** combinations, showcasing where each model achieved its most accurate predictions.

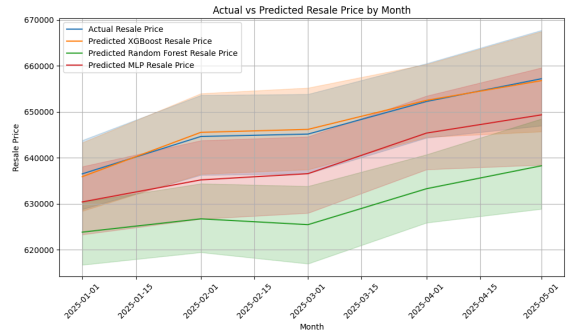
Actual vs Predicted Resale Price:

Implementation 1: Without Feature Engineering



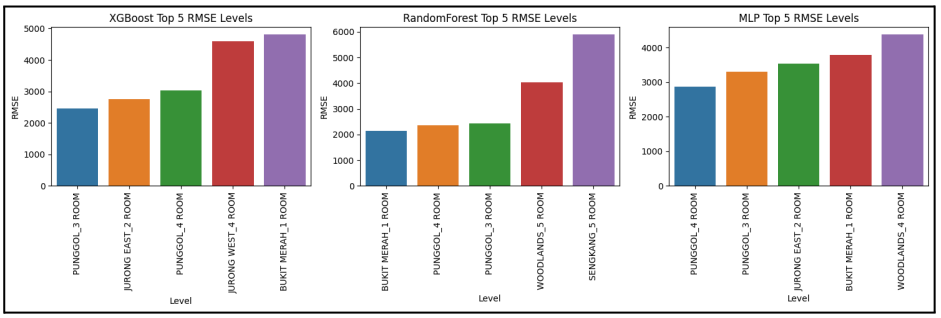
Implementation 1, *all models predict lower than the actual prices*. XGBoost and Random Forest follow the upward trend but stay below the actual values. MLP performs the worst, with flat predictions that miss the trend completely. Overall, the models show weak performance and low accuracy.

Implementation 2: Feature Engineering



Implementation 2, *the models perform much closer to the actual price*. XGBoost gives the most accurate results, closely matching actual prices. Random Forest improves and follows the trend well, though it sometimes predicts slightly too high. MLP also improves but is still less accurate than the others. This version shows stronger model performance and better tracking of real market changes.

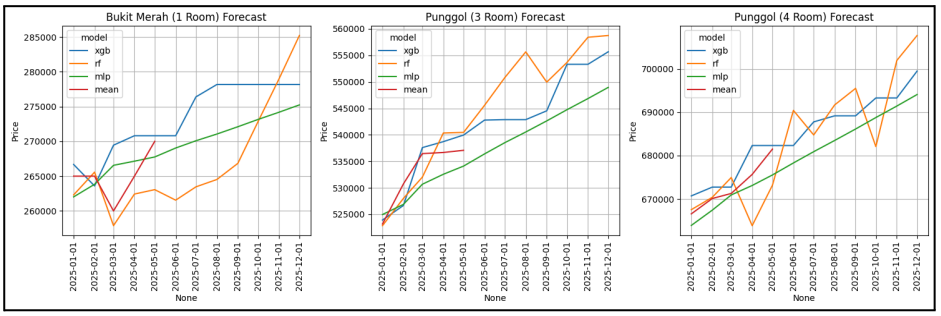
3.6 SKForecast



Best RMSE (skforecast)

The lowest RMSE values were observed in flats located in **Punggol**, **Jurong East**, and **Bukit Merah**, particularly for smaller unit types like 1–4 Room flats. **XGBoost** consistently showed the smallest RMSE values, indicating strong accuracy in stable housing markets. While **Random Forest** and **MLP** performed comparably, both had slightly higher errors, especially in larger or more variable flat types.

Backtest Metrics (skforecast)



In **Bukit Merah (1 Room)**, all models followed the trend well, but **XGBoost** offered the most consistent tracking close to actual prices. **Punggol (3 Room and 4 Room)** forecasts also favored XGBoost and Random Forest, which captured monthly changes more accurately than MLP. MLP trended more conservatively across all plots, suggesting weaker adaptability to rapid market shifts.

4.0 Conclusion

While our models achieved strong performance in predicting HDB resale prices, there are clear areas for future improvement. One major limitation was the imbalance in transaction data across different towns and flat types. High-volume areas tended to dominate the learning process, which led to reduced accuracy in less common cases such as multi-generation flats or premium units. To mitigate this, future iterations of the model should explore techniques like oversampling underrepresented categories or using weighted loss functions to ensure more balanced learning.

Another area of opportunity lies in the inclusion of geospatial data. The current dataset does not account for location-specific influences such as distance to MRT stations, schools, or other amenities, which are often key factors in real-world pricing. Adding this layer of spatial context could significantly improve the model's ability to capture buyer behavior and local demand. Furthermore, incorporating exogenous variables such as inflation rates, policy shifts, or development activity may allow the model to respond more accurately to macroeconomic changes and market volatility.

Looking ahead, adopting more flexible and adaptive architectures—such as ensemble or hybrid models—can help enhance generalization across a wider variety of flat types and town profiles. With these enhancements, the model can evolve into a more robust, practical, and equitable tool for forecasting public housing prices in Singapore.

5.0 References

Brownlee, J 2020, 'How to Develop Multilayer Perceptron Models for Time Series Forecasting', 28 August, viewed 18 May 2025, <https://machinelearningmastery.com/how-to-develop-multilayer-perceptron-models-for-time-series-forecasting/>.

Jaiswal, S 2025, Multilayer Perceptrons in Machine Learning: A Comprehensive Guide, 5 April, viewed 19 May 2025, <https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning>.

Matalonga, H 2023, Choosing between MAE, MSE and RMSE, viewed 18 May 2025, <https://hmatalonga.com/blog/choosing-between-mae-mse-and-rmse/>

Michael Affenzeller, Aleksandra Petrakova, & Galina Merkuyeva 2015, 'Heterogeneous versus Homogeneous Machine Learning Ensembles', *Heterogeneous versus Homogeneous Machine Learning Ensembles*, December, viewed 11 May 2025, <https://www.researchgate.net/publication/293194221_Heterogeneous_versus_Homogeneous_Machine_Learning_Ensembles>.

Michael Affenzeller 2015, *The common ensemble architecture fig1*, viewed 11 May 2025, <https://www.researchgate.net/figure/The-common-ensemble-architecture_fig1_293194221>

Quang Truong, Minh Nguyen, Hy Dang, & Bo Mei 2020, *Housing Price Prediction via Improved Machine Learning Techniques*, 27 July, viewed 15 May 2025, <https://doi.org/10.1016/j.procs.2020.06.111>.

Rostami, J 2019, *Time Series Forecasting of House Prices: An evaluation of a Support Vector Machine and a Recurrent Neural Network with LSTM cells*, 17 June, viewed 5 May 2025, <https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-385823>.

Skforecast 2024, User guides – Autoregressive forecaster, viewed 18 May 2025, <https://skforecast.org/0.16.0/user_guides/autoregresive-forecaster.html>.

Skforecast 2024., *Time series differentiation*, viewed 12 May 2025, <https://skforecast.org/0.16.0/user_guides/time-series-differentiation.>.

Wan Teng Lim, Lipo Wang, Yaoli Wang, & Qing Chang 2016, 'Housing price prediction using neural networks', *Housing price prediction using neural networks*, 24 October, viewed 12 May 2025, <https://doi.org/10.1109/FSKD.2016.7603227>.

WY 2022, 'Choropleth Mapping Singapore Using Python', *Choropleth Mapping Singapore Using Python*, 20 June, <https://medium.com/@lwyeong/choropleth-mapping-singapore-using-python-24cb26173fdd>.

6.0 Appendix

GitHub Repository

https://github.com/lester-liam/csci323-housing-price-prediction

Google Colab Notebooks

The Google Drive Link Folder Provide View Access to our Colab Notebooks with cell outputs:
https://drive.google.com/drive/folders/1Ez3SbD3WR1AzYgfoJXDtC7smDhS7tnyc?usp=sharing

Acknowledgement: Table of Contributions

S/N	Student's Name	Contribution	%
1	Lester Liam Chong Bin	Project lead; handled model evaluation, RMSE/MAE comparison, and visualizations.	16.66
2	Bryce Nicolas Fernandez Sumcad	Implemented and analyzed Skforecast (time series), performed EDA, and wrote time series discussion and conclusion.	16.66
3	Jesyln Ho Ka Yan	Handled theoretical research, literature review, and drafted background theory and model explanation sections.	16.66
4	Park Ki Sung	Model Builder – supported MLP and Skforecast implementation, ran feature-engineering tests, and assisted with graphs.	16.66
5	Lee Donghyun	Model Builder – assisted with model implementation, testing, and validation across towns and flat types.	16.66
6	Chea Darayuth	Built Random Forest and XGBoost models, tuned hyperparameters, and integrated results into report and slides.	16.66