Predicting Prices of Private Residential Properties in Singapore

Team 8 Members: Wong Ya Chong Jerome, Mehmet Balkan, Nikoloz Jaghiashvili

1.0 Introduction

This study aims to understand and predict private condominium prices in Singapore using five years of transactional data (2020 to 2025). This will assist potential buyers in identifying key factors influencing property prices and support them in their decision-making process. The motivation for this study arises from the relative lack of studies focused specifically on private condominiums in Singapore. To establish a clearer understanding of the data, we first performed outlier detection and clustering. Outlier detection uncovered anomalous bulk sales, which were excluded from subsequent supervised learning, while clustering revealed natural groupings of transactions based on location. Supervised learning results showed that unit area had a substantially greater influence on prices compared to other features.

2.0 Related Work

Several studies have investigated the use of predictive models to estimate property prices in Singapore. Previously, Tay (2024)[1.] developed a model to predict the prices of private condominiums in Singapore. While the objective is similar, our project builds on it by incorporating unsupervised learning techniques and additional features from secondary data sources. Additionally, previous studies such as Calis (2023)[2.] and Mallaiyan (2024)[3.] have focused on predicting the prices of public resale flats. In contrast, our project is exclusively concerned with predicting the prices of private condominiums. Furthermore, whereas Calis (2023) limited the scope of his work to Linear Regression, our study extends this by experimenting with Random Forests and Support Vector Regression.

3.0 Data Source

The following datasets are employed for both supervised and unsupervised learning. A list of features used in the modeling process can be found in Appendix A.

Table 1: Primary Data

Source	Description	Important Features	Record Count	Format
Urban Redevelopment Authority (URA) API	Private Residential Transactions (2020 to 2025)	1. project 2. district 3. marketSegment 4. x, y (coordinates) 5. area 6. contractDate 7. tenure 8. price	141,332	JSON

Table 2: Secondary Data (Amenities)

Source	Description	Important Features	Record Count	Format
Land Transport Authority (LTA)	Train Stations	mrt_stations_english	213	xls
OneMap API	Train Stations Coordinates	x, y (coordinates)	213	JSON
Google Places API	Points of Interest	1. name 2. price_level 3. user_ratings_total 4. rating 5. Vicinity 6. x, y (coordinates) 7. poi_type 8. fetch_date 9. delta_rating_count 10. delta_time	25,272	pickle

Table 3: Secondary Data (Economic Indicators)

Source	Name	Important Features	Record Count	Format
Monetary Authority of Singapore (MAS)	Singapore Overnight Rate Average (SORA)	SORA	251	csv
Singapore Department of Statistics(Singstat)	Consumer Price Index	1. cpi 2. cpi_accumulated 3. target_price_cpi_adjusted	772	xlsx
Singapore Department of Statistics(Singstat)	Population Growth	monthly_population_growth_rate	75	xlsx
Singapore Department of Statistics(Singstat)	Marriage Rates	monthly_marriage_crude_rate	44	xlsx
Singapore Government Agency(Data.gov)	Private Home Index	monthly_price_index	603	csv

4.0 Feature Engineering

Feature engineering enriches the model by incorporating domain knowledge and external data. Information on amenities and points of interests were integrated to enrich the dataset with location specific context. Additionally, economic indicators are used to reflect macro level demand, affordability together with market sentiment.

4.1 Primary Dataset

To account for the temporal component in the dataset, a feature was created to represent the number of days since the first recorded transaction. In addition, the original tenure feature (e.g., '99-year lease commencing from 2007') was used to derive both the lease and the age of the property. The X and Y coordinates (SVY21) were also converted to latitude and longitude (WGS84) to support integration with external spatial data sources.

4.2 Distance to Train Stations

The names of train stations were first obtained from LTA and subsequently categorized into Mass Rapid Transit (MRT) and Light Rail Transit (LRT) stations. The MRT system is designed for long-distance travel, connecting key regions across Singapore, whereas the LRT serves shorter routes and functions as a feeder network to the MRT. Following which, the OneMap API was used to extract the latitude and longitude of each station, and GeoPandas was employed to compute the distance from each private condominium to the nearest train station.

4.3 Google Points of Interest (POI)

To enhance the location intelligence of our dataset, we integrated data from the Google Places API, a robust service that provides detailed information about POIs around a given location. By leveraging this API, we were able to retrieve structured data on nearby amenities, including restaurants, hotels, shopping malls, hospitals, schools, and police stations. The API provides up-to-date reviews and descriptions for each POI.

The Google Places API allows querying POI data within a specified spatial radius, returning up to 60 results per request. To address this limitation and ensure comprehensive spatial coverage, we utilised a polygon shapefile defining Singapore's administrative boundary (sourced from data.gov.sg). Within this boundary, a grid of overlapping circular buffers with a 550-meter radius was generated. These buffers acted as sampling zones for API queries. The relatively small radius was chosen to reduce the risk of exceeding the 60 POI limit per query, thereby allowing for a more accurate and granular mapping of specific POI types across the entire area (refer to Appendix B for illustration).

The output of the Google Places API querying process is a long table, where each row corresponds to a single point of interest (POI),with metadata such as type, name, user rating, and geographic coordinates. In the subsequent feature engineering stage, this amenity dataset is spatially joined to the core dataset.

Specifically, for each property transaction, POIs within a 500-meter buffer radius are obtained using the sjoin function from the GeoPandas library. The matched POIs are then aggregated and pivoted to generate type-specific features, such as the count, average rating. Properties with no matches are imputed with 0 to indicate the absence of surrounding POIs.This transformation results in a wide-format feature table that captures the amenities around each property, which is then used as input for modelling tasks.

Google Places API does not provide business inception dates, posing a risk of data leakage given the five-year core dataset. To mitigate this, POI data was queried at two time points to calculate changes in review count (delta_rating_count) over time (delta_time), to predict business inception date. However, due to the short interval between queries, this method proved ineffective. We still include it as a potential avenue for future improvement, further discussed in the report's Discussion section.

The feature engineering pipeline included the following steps:

1. **Query POI Data via Google Places API** – Queried POIs using a grid of overlapping (550m radius and 700m between centroids) circles across Singapore's administrative boundary.
2. **Extract POI Metadata** – Parsed key attributes from raw API responses: name, type, coordinates, user rating, and total number of reviews.
3. **Query POI Data Second time**– Queried POI data at two points in time to estimate business age via change in review count (delta_rating_count). This approach was not feasible due to the limited time window.
4. **Spatial Join with Core Dataset** – Used GeoPandas's sjoin with a 500m buffer to associate each transaction with nearby POI.
5. **Pivot POI Types into Features** – Aggregated POIs by type for each property to create wide-format features (num_poi_count_restaurant, num_avg_rating_restaurant, num_avg_price_level_restaurant etc.).

4.4 Economic Indicators

To incorporate broader market dynamics and account for changing economic and demographic conditions over time, we included variables that serve as proxies for overall wealth, purchasing power, housing affordability, and household formation trends. These help us better capture demand and supply dynamics beyond project-level attributes.

Table 4: Feature Engineering For Economic Indicators

Source	Name	Purpose and Transformation
Monetary Authority of Singapore (MAS)	Singapore Overnight Rate Average (SORA)	Proxy to reflect borrowing cost and mortgage affordability. Used month end values as monthly average rate to reflect financing conditions.
Singapore Department of Statistics(Singstat)	Consumer Price Index	Reflects price appreciation / depreciation over time. Used housing and utilities sub index , computed monthly changes. And calculated nominal house prices excluding CPI.
Singapore Department of Statistics(Singstat)	Population Growth	Proxy for housing demand pressure. Derived monthly growth rate from yearly figures and estimated growth using linear regression for missing months.
Singapore Department of Statistics(Singstat)	Marriage Rates	Represents household formation. Derived monthly growth rate from yearly figures and estimated growth using linear regression for missing months.
Singapore Government Agency(Data.gov)	Private Home Index	Key benchmark indicator of overall private home prices and market trends, Calculated monthly rate from compounded quarterly indices.

5.0 Unsupervised Learning

5.1 Outlier Detection

5.1.1 Methods Description

Motivation
The outlier detection model was applied to the entire dataset, specifically on target_price, area, noOfUnits, and MarketSegment as these are key property attributes. The motivation is to detect extreme values that can bias the results of the supervised learning process.

Data Preprocessing
Before performing outlier detection, categorical features were one-hot encoded, while numerical features were scaled using min-max normalization. This ensures that the data is suitable for modeling and prevents features with larger scales from dominating the results. For neural networks, normalizing the data also enables optimization algorithms such as gradient descent to converge faster during the training process.

Model Selection
For outlier detection, two models with different underlying mechanisms were employed: 1) Isolation Forest and 2) Autoencoder. The Isolation Forest and Autoencoder were selected as they are better equipped to handle mixed data compared to distance-based algorithms such as k-Nearest Neighbors (kNN), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Local Outlier Factor (LOF). Additionally, the Isolation Forest is scalable and efficient, achieving linear complexity, while the Autoencoder, as a neural network, is able to learn complex non-linear relationships within the data.

Hyperparameters

a. Isolation Forest
The n_estimators represent the number of trees in the ensemble. In general, using more trees leads to more stable and accurate results, but it also increases the computational time and memory required to train the model. The n_estimators were set to 50, 100, 200 and 300 to assess how the number of trees affect outlier detection performance.

Table 5: Isolation Forest Hyperparameters Settings

Name	n_estimators	max_features	Default Setting
IForest 1	50	1	No
IForest 2	100	1	Yes
IForest 3	200	1	No
IForest 4	300	1	No

b. Autoencoder
The hidden_neuron_list represents the number of neurons in each layer of the Autoencoder, controlling how the data is compressed and reconstructed (e.g., [64, 32] creates an encoder of [64, 32] and a decoder of [32, 64]). The epoch_num represents the number of iterations during which the entire dataset is passed through the model. In general, increasing the number of neurons, layers, and training epochs can help the model capture more complex data.

Table 6: Autoencoder Hyperparameters Settings

Name	hidden_neuron_list	epoch_num	Default Setting
Autoencoder 1	[64, 32]	10	Yes

Autoencoder 2	[64, 32]	20	No
Autoencoder 3	[128, 64, 32]	10	No
Autoencoder 4	[128, 64, 32]	20	No

5.1.2 Model Evaluation

With sufficient labels, traditional metrics such as Precision, Recall and F-1 scores can be used to assess the performance of outlier detection models. However, in the absence of labels, visualization charts were used to estimate model performance. To enable clearer visualization, the dataset was downsampled by selecting the top 20 property transactions with the highest outlier scores, along with a random sample of 200 records from the rest of the data. Next, a Gower distance matrix was computed to handle both numerical and categorical features.This matrix was then used to generate a two-dimensional scatter plot using Uniform Manifold Approximation and Projection (UMAP). The outliers are colored in red while the inliners are colored in blue. Based on the UMAP visualizations, the best model was Isolation Forest 3 as it showed the clearest separation between outliers and inliers.

Table 7: Best Model

Model Family	Best Model	Overall Best
IForest	IForest 3	Yes
Autoencoder	Autoencoder 4	No

Figure 1: Best Isolation Forest UMAP

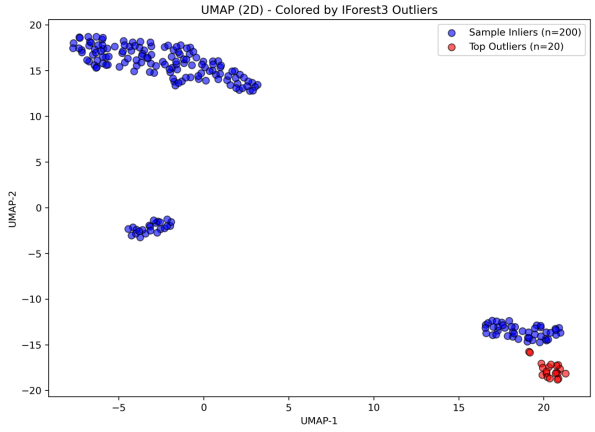
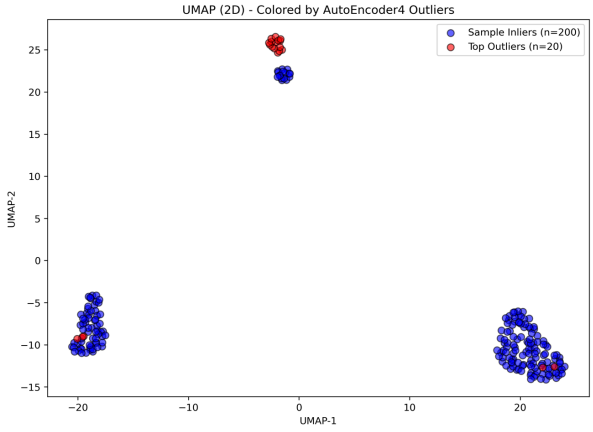


Figure 2: Best Autoencoder UMAP



Sensitivity Analysis

The max_features parameter determines the maximum number of features to use when training each base estimator. It was observed that the results slightly degraded when the max_features was set to 4 (refer to Appendix C). This might be because there were only four features, and using all of them in every tree could have reduced the diversity of the ensemble, weakening its overall effectiveness. As there were no discernible differences among the remaining models, the best model settings remained at n_estimators = 200 and max_features = 1.

Table 8: Sensitivity Analysis Best Model Settings

Name	n_estimators	max_features	Best Setting
IForest 31	200	1	Yes
IForest 32	200	2	No
IForest 33	200	3	No
IForest 34	200	4	No

5.2 K-Means Clustering

5.2.1 Methods Description

The objective of this task is to segment condominiums and identify distinct property clusters based on multidimensional numerical features, providing insights for modelling and uncovering hidden market structures and associated characteristics. This approach allows us to move beyond human-defined categories (e.g. districts, typeOfSale) and enable insights to emerge directly from the data. The features are selected based on following criteria:

Table 9: K-Mean Clustering Features

Feature	Justification
area	Differentiates unit preferences based on demographic choices and pricing tiers, such as family-oriented projects with larger units or smaller units in high-density areas.
lrt_nearest_distance_m mrt_nearest_distance_m	Proximity to public transit significantly influences condominium developments, making it one of the most important factors for buyers.
poi_count_shopping_mall	Projects near malls are often seen as attractive options and tend to cater to a retail-centric lifestyle.
poi_count_restaurant	Areas with many eateries are more appealing to working professionals and younger generations, making it a key factor driving buyer demand.
poi_count_police	Captures safety infrastructure and defines major areas for project developers to consider.
poi_count_school	Proximity to schools is a key determinant for families when choosing a home. In Singapore, priority for primary school registration is granted based on distance, and many families prefer to avoid long commutes for their children.
tenure_bin	Affects the long-term value, mortgage approval, and pricing of units. Freehold and 999-year leasehold units tend to have higher value compared to those with a 99-year lease. Additionally, mortgage approval becomes more difficult once a project reaches 30 to 40 years of age.

Categorical features such as district, typeOfSale and marketSegments were excluded as they are not meaningful as one-hot encoded categorical variables. These features can distort distance by introducing artificial similarity or dissimilarity. It also can dilute influence of important numerical features and increase risk of dimensionality.

Numerical features that were excluded include poi_count_lodging and poi_count_hospital. Lodging is primarily located in prime areas such as financial hubs and tourist zones like Orchard and Marina Bay. These areas reflect economic clusters rather than typical residential preferences. To avoid bias toward commercial zones, lodging was excluded from clustering. Also, average ratings features were excluded as they contain missing data.

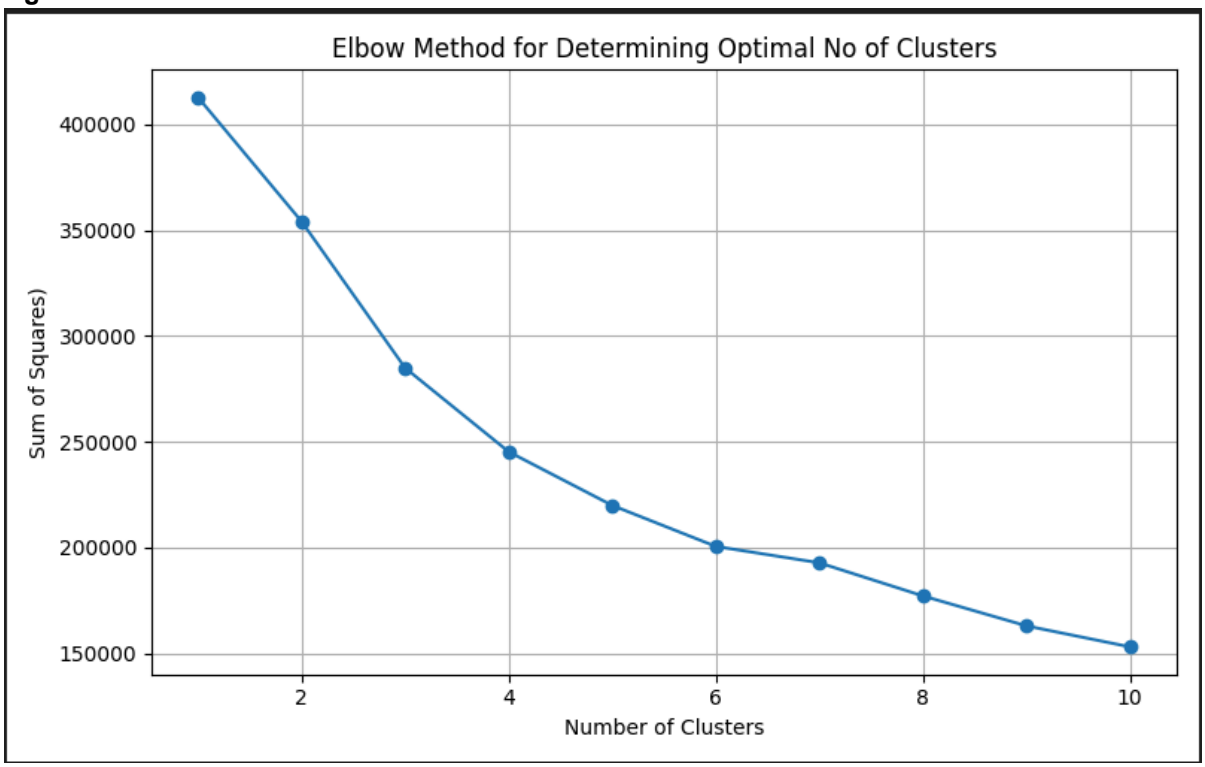
5.2.2 Model Evaluation

Determining the Number of Clusters

Elbow Method

The elbow method is used to determine the number of clusters based on the sum of squared errors (SSE). It shows an elbow at k = 4, indicating diminishing returns beyond this point. Adding more clusters after 4 results in minimal reduction in variance, suggesting that 4 is the optimal choice from a variance reduction perspective.

Figure 3: Elbow Method

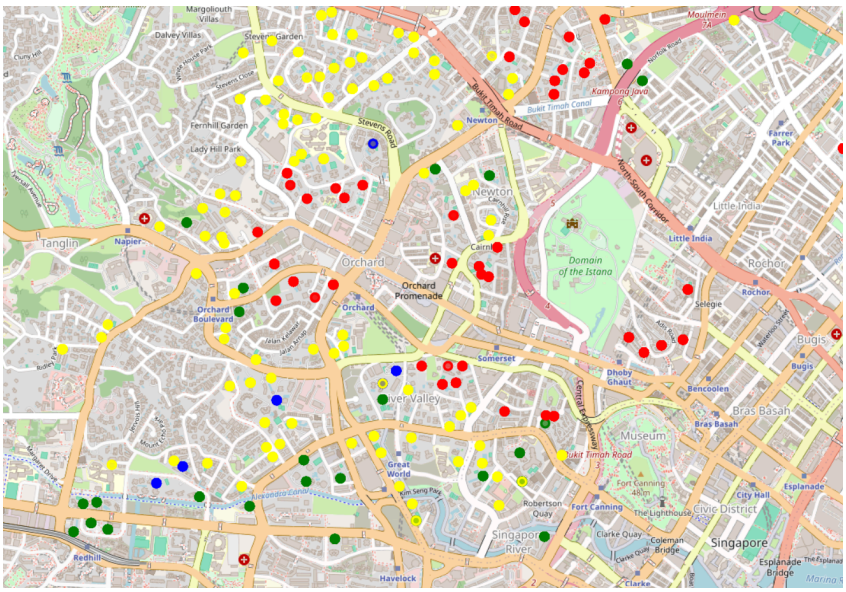
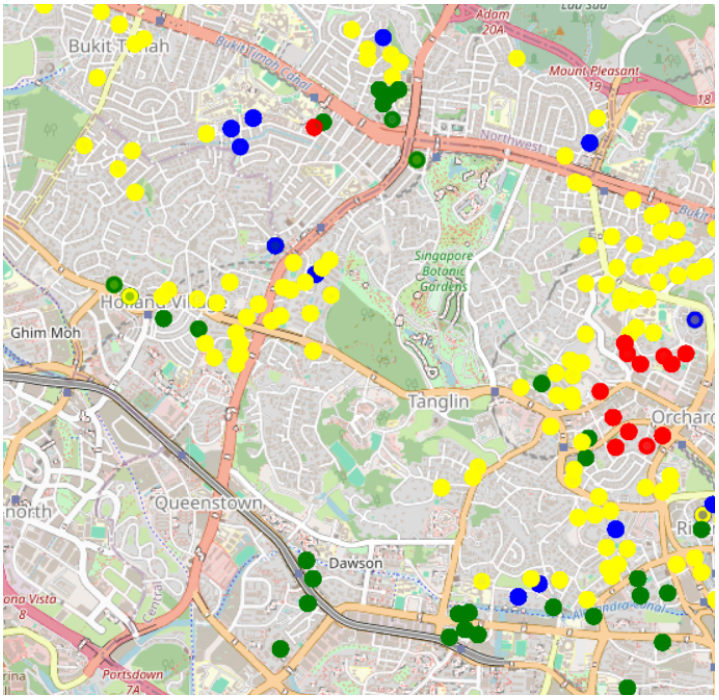


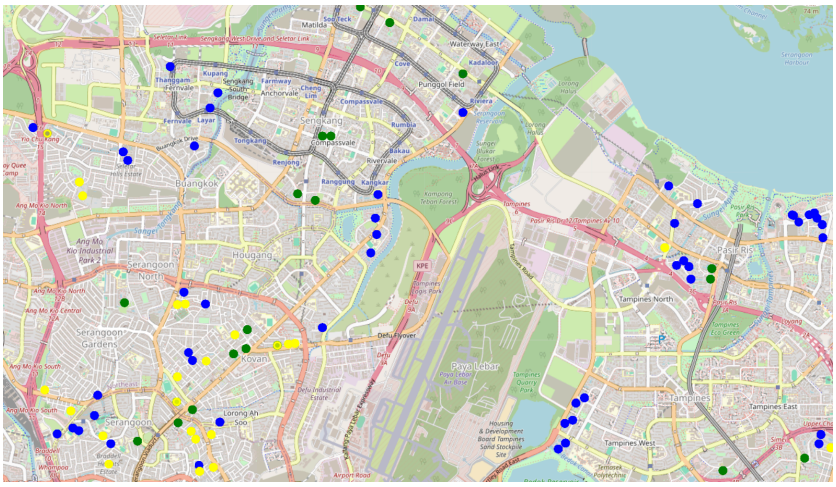
Silhouette Score
The silhouette score is 0.2494 when the cluster count is 4. Typically, a score between 0.2 and 0.3 indicates meaningful structure, making it suitable for exploratory market segmentation in the housing market domain. Therefore, both the Elbow method and silhouette score indicate that 4 is the ideal number of clusters.

Visualizing the Clustering Results

Cluster Visualization Map
This allows intuitive interpretation of the housing clusters and shows Singapore characteristics and identifies location driven trends. It focuses on numerical data, ignoring human-defined segmentations.

Table 10: Cluster Visualization Map Results

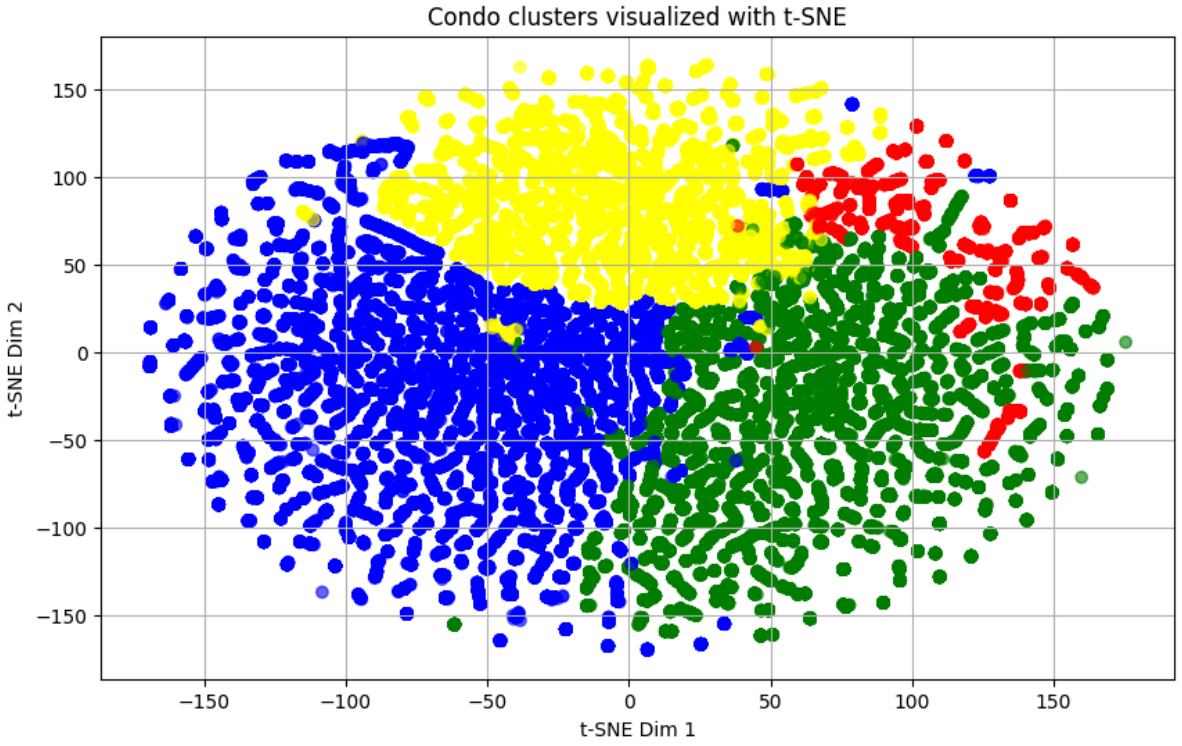
Cluster 1 A dense concentration of red points is observed in central areas such as Orchard, Newton, River Valley, and Marina Bay. These prime districts (Districts 9, 10, and 11) fall within the Core Central Region. This cluster likely represents luxury condominiums and reflects premium pricing.	
Cluster 2 and 3 The yellow and green clusters are found in areas such as Bukit Timah, Queenstown, Bishan, Marine Parade, Bedok, and Clementi. These neighborhoods cater to mid-tier buyers, offering a balance of affordability and access to amenities like schools and MRT stations.	

Cluster 4 Suburban areas such as Jurong, Woodlands, Punggol, Sengkang, Changi, and Pasir Ris show a concentration of blue points. These clusters primarily consist of larger, family-sized units, aligning with the government's decentralization strategy.	

t-SNE Plot

The t-SNE plot reduces high-dimensional features to two dimensions, effectively revealing distinct and separable cluster structures. This technique preserves local neighborhoods, making it effective for visualizing the complex, non-linear relationships present in features such as proximity to transport and tenure type. Each dot in the plot below represents a sales transaction. Although there are slight overlaps due to transitional characteristics, the plot reveals meaningful segmentation patterns. The cluster colors have been standardized to align with those in the previous spatial map, facilitating easier comparison and reinforcing the insights.

Figure 4: t-SNE Plot



6.0 Supervised Learning

6.1 Data Splitting
The data was divided into a training set and a held-out test set based on contract date, ensuring that test records occurred after those in the training set. The training data was then subjected to 5-fold cross-validation. Within each fold, the data was further split into training and validation subsets in chronological order.

6.2 Data Preprocessing
In each fold, missing numerical values in points-of-interest features were imputed with zeros to reflect the absence of nearby amenities. Categorical features were then one-hot encoded, and numerical features were scaled using min-max normalization. This ensures the data is suitable for modeling and prevents features with larger ranges from dominating the results.

6.3 Model Selection
Since the target variable is numerical, models from three distinct regressor families were selected to predict price: (1) Linear Regression, (2) Random Forest Regressor (RFR), and (3) Support Vector Regression (SVR). Linear Regression was selected as a baseline, whereas RFR and SVR were included as more advanced models to capture complex non-linear relationships in the data. Additionally, the RFR was chosen for its robustness to overfitting due to its ensemble nature, while the SVR performs well on both low- and high-dimensional datasets.

6.4 Hyperparameters
The max_depth parameter in the Random Forest Regressor controls how deep each tree can grow. In contrast, the C parameter in SVR controls the degree of regularization, with larger values of C corresponding to less regularization.To explore their effects, max_depth was set to 5, 10, and 15, while C was set to 0.1, 0.5, and 1.0. A grid search was then conducted to evaluate the performance of each hyperparameter combination.

Table 11: Supervised Learning Hyperparameters Settings

Name	Hyperparameters	Settings
Linear Regression	All	Default
RFR	max_depth	[5, 10, 15]
SVR	C	[0.1, 0.5, 1.0]

6.5 Model Results
Mean Absolute Error (MAE) was used to determine the best model. It represents the average absolute difference between the predicted and actual prices. MAE was chosen because it is more interpretable than other evaluation metrics, such as R-squared or Root Mean Squared Error (RMSE), since it is measured in the same units as the target variable. Based on the mean MAE obtained from 5-fold cross-validation, the overall best model is RFR 3, which achieved the lowest average MAE of S$155,281. This is 54.9% lower than the Linear Regression baseline. Among the SVR models, SVR 3 is the best performer, although its MAE was substantially higher at S$797,736.

Table 12: Supervised Learning Model Results

Name	Hyperparameters	Mean MAE 5-Fold CV	Standard Deviation	Best in Family	Overall Best
Linear Regression	Default	S$344,238	±S$26,924	Yes	No
RFR 1	max_depth = 5	S$337,060	±S$23,604	No	No
RFR 2	max_depth = 10	S$199,630	±S$28,151	No	No
RFR 3	max_depth = 15	S$155,281	±S$11,405	Yes	Yes
SVR 1	C = 0.1	S$797,990	±S$47,815	No	No
SVR 2	C = 0.5	S$797,876	±S$47,753	No	No

SVR 3	C = 1.0	S$797,736	±S$47,676	Yes	No

6.6 Feature Importance and Ablation Analysis

6.6.1 Feature Importance Analysis
With a feature importance score of 0.63, area outperforms all other features in the feature importance analysis of the best performing model. It is the most significant factor in predicting private condominium prices. This is followed by marketSegment features (e.g. Core Central Region, Out of Central Region). These characteristics demonstrate the model's sensitivity to location and premium zones by capturing the regional and prestige-based stratification of properties.

Other important features that affect buyer preference are proximity to public transportation as indicated by lrt_nearest_distance_m and mrt_nearest_distance_m. It was also interesting to note that local restaurants and hospitals have significant influence, reflecting the perceived quality of surrounding facilities. Overall analysis confirms that both physical and environmental attributes shape the true value of condominiums in Singapore's private housing market.

Table 13: Top 10 Feature Importances

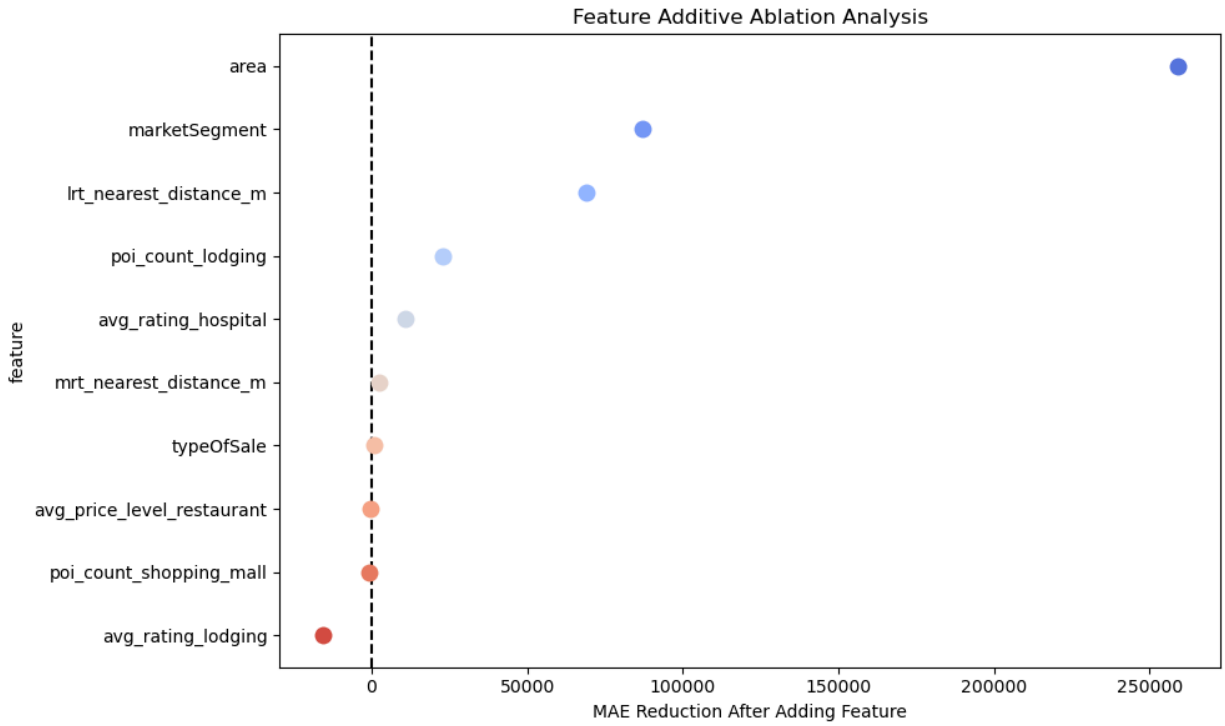
Rank	Feature	Importance
1	num__area	0.63
2	cat__marketSegment_CCR	0.062
3	cat__marketSegment_OCR	0.048
4	num__lrt_nearest_distance_m	0.043
5	num__poi_count_lodging	0.021
6	num__avg_price_level_restaurant	0.016
7	num__avg_rating_lodging	0.015
8	num__avg_rating_hospital	0.015
9	num__mrt_nearest_distance_m	0.014
10	cat__typeOfSale_3	0.013

6.6.2 Additive Feature Ablation Analysis
An additive ablation analysis was performed by incrementally adding each of the top 10 features to a minimal baseline model, measuring their contribution to reducing the MAE.

Overall, area resulted in the greatest MAE reduction, indicating that property size is the most influential driver of price. Next, transport accessibility, captured by lrt_nearest_distance_m and mrt_nearest_distance_m, emerged as another key factor. Properties located closer to public transit options tended to have higher prices, reflecting buyers' preference for convenience and connectivity.

Market related categories like marketSegment and typeofSale also contributed significantly to predicting prices, followed by lodging, hospital and restaurant related features. This stepwise additive approach provided insight for the marginal value of each feature, and aided feature selection and prioritization during model development.

Figure 5: Feature Additive Ablation Analysis



6.6.3 Feature Set Ablation Analysis

To better understand group-level contributions, a feature ablation analysis was conducted by dividing the features into three sets: 1) Primary Data, 2) Amenities and POIs and 3) Economic indicators. The results showed that primary features achieved the lowest MAE, followed by amenities and economic indicators.

Table 14: Feature Set Ablation Analysis

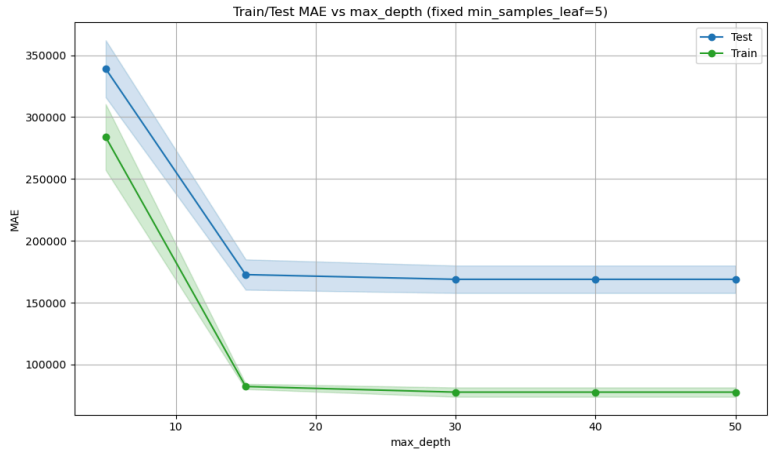
SN	Feature Set	Mean MAE 5-Fold Cross-Validation	Standard Deviation
1	Primary Features	S$195,461	±S$12,839
2	Amenities and POIs	S$483,736	±S$34,563
3	Economic Indicators	S$821,744	±S$43,946

6.7 Sensitivity Analysis

Hyperparameter sensitivity analysis was performed to identify which hyperparameters the model is most sensitive to and to assess the model's stability with respect to changes in those parameters. Additionally, sensitivity analysis provides insight into the model's ability to generalise. By evaluating performance on both training and validation sets during cross-validation, this analysis goes beyond simple grid search by revealing the degree of overfitting and the variability in performance across the hyperparameter grid.

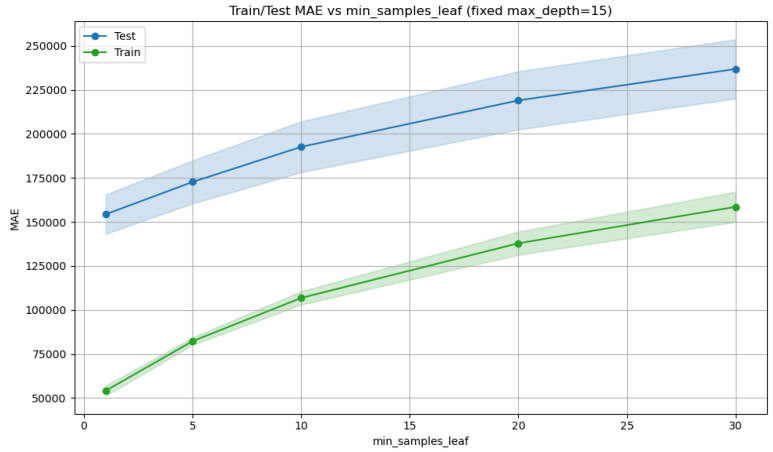
During the analysis, we identified max_depth and min_samples_leaf as the most influential hyperparameters for our top-performing Random Forest model. We first performed sensitivity analysis by varying each hyperparameter independently, and then examined their interaction using a 3D performance surface plot. Sensitivity was evaluated separately on the training and validation sets using 5-fold time series cross-validation, allowing us to assess both overfitting and generalisation behaviour across the hyperparameter space.

Figure 6: Supervised Learning Sensitivity Analysis



1. Our Random Forest model is quite sensitive to the max_depth parameter. As seen on the first plot, the MAE drops sharply as max_depth increases from 5 to 15, beyond which the metric stabilises and is consistently flat up until max_depth of 50. The model exhibits a notable gap between the validation and train sets, which increases by a factor of two as max_depth increases from 5 to 15 and stabilises afterwards. The overfitting is present even when performing a sensitivity analysis on a model trained on just core features, a possible reason for which is discussed later in section 6.9. The metric variability is quite low in both test and train sets, it drops consistently with the MAE value and then stabilises beyond a max_depth of 15. Based on these observations, the initial choice of max_depth of 15 is well justified.

Figure 7: Supervised Learning Sensitivity Analysis



2. The min_samples_leaf parameter is introduced to reduce overfitting by enforcing a minimum number of samples required to form a leaf node. However, increasing the hyperparameter does little to alleviate the degree of overfitting observed above. As min_samples_leaf, the MAE increases together with its variability. The difference between train and validation performance is consistent across different values, which also indicates that leaving the default min_samples_leaf for the best model is justified.

The interaction between the two hyperparameters does not offer any additional insights beyond what was observed in the individual analyses. The corresponding plot can be found in Appendix F.

6.8 Failure Analysis

The top ten records in the held-out test set with the greatest absolute errors were isolated for failure analysis. Among these, three broad categories of failure were uncovered from three specific instances.

Table 15: Categories and Instances of Failure

Row No.	Category	Project Name	Actual Price	Predicted Price	Absolute Error	Direction
4,576	Systematic Error	Turquoise	S$9.0m	S$18.8m	S$9.8m	Overpredict
10,137	Edge Cases	21 Anderson	S$24.0m	S$11.0m	S$13.0m	Underpredict
5,411	Random Error	Sophia	S$4.8m	S$7.9m	S$3.1m	Overpredict

6.8.1 Systematic Errors
The model tends to predict exceptionally high prices for large private condominiums exceeding 600 square meters. In the absence of other informative predictors, it may be over-relying on the area feature to estimate property prices. In the case of Project Turquoise, the SHAP waterfall plot (refer to Appendix D) shows that area alone contributes S$14.1 million to the final prediction of S$18.8 million. However, the actual price of the unit was only S$9.0 million.

6.8.2 Edge Cases
Additionally, some unusual cases in the data have a disproportionate impact on the MAE. For example, units at Project 21 Anderson were sold for more than S$20 million - far surpassing prices of similarly sized properties in the Core Central Region. The development commands such high prices because it is located within an elite enclave and specifically targets the ultra-rich.

6.8.3 Random Errors
Finally, there are random errors that do not follow a clear pattern. These may be one-off cases in the dataset. For instance, the model predicted S$7.9 million for a unit at Sophia Residence, while a nearly identical unit was estimated at only S$3.5 million. Such large discrepancies between similar units suggest that the model may be sensitive to minor variations in the data.

6.8.4 Future Improvements
To address the aforementioned issues, it may be necessary to include additional features such as the developer, since properties from well-known developers can command higher prices regardless of their area. Applying a log-transformation to the area feature can also help reduce the influence of extremely large units in the data. Furthermore, investigating the use of neural networks may enhance the modeling of complex relationships and improve predictive accuracy.

6.9 Trade-Offs
The consistent degree of overfitting in the model is most likely due to a lack of apartment-specific features. The high degree of accuracy is a significant advantage of the transaction dataset, while a lack of apartment-level features, like layout, renovation, and interior design, is a significant drawback. The missing features create a sharp trade-off between model performance and overfitting, which mainly occurs when tuning max_depth between values 5 and 15. Based on the relevant literature, interior and layout variables can marginally improve model R2 by between 5-10%. Leung, Ma, and Zhang (2014) demonstrated, based on the Hong Kong real estate market, that incorporating the net_ration variable (ratio of livable space in the apartment) results in a 5.6% improvement in R2, while further adding interior and layout variables adds 2%. Mamre and Sommervoll (2024) quantified, based on the Norwegian real estate market, a 5–7% price premium for renovated homes.

7.0 Discussion

7.1 Google POI Data
To mitigate potential data leakage in Google POI features, we developed a separate supervised learning algorithm to predict business inception dates. The independent variables included the change in the number of reviews and other descriptive features. The target variable was derived by matching Google Business names with the ACRA corporate entity database, which contains over 2 million registered entities with known inception dates. Using fuzzy matching on both business names and addresses, we identified approximately 700 entities with matching scores above 90. Despite the successful matches, the two-month time difference did not provide sufficient explanatory power to train a reliable model. However, extending the time window could potentially improve model performance. The code for this exercise was included in the repository.

7.2 Unsupervised Learning

a. Interestingly, the outlier detection model uncovered bulk sales — transactions of multiple units — which were not apparent in the original dataset. These were excluded from the supervised learning process, as they disproportionately affect the target price and area.
b. Without labels, a key challenge was assessing model performance, as traditional metrics such as Precision, Recall, or F1-scores could not be applied. To address this, the outliers were visually inspected using dimensional reduction techniques such as UMAP.
c. Instead of relying on a single outlier detection model, the solution could be extended in the future by ensembling the results from multiple detectors to achieve more robust and reliable detection.

7.3 Supervised Learning

a. Based on the feature importance scores, it was surprising that a property's area had a much higher impact compared to other features, even surpassing district or market segment (e.g. Core Central Region, Rest of Central Region and Outside Central Region). As a follow-up, it would be worthwhile to further investigate this phenomenon.
b. A primary challenge in this task was developing a customized cross-validation pipeline such that transactions in the validation data occurred chronologically after the train data. In response, we utilized the TimeSeriesSplit module from Scikit-Learn to construct it.
c. In the future, an interactive choropleth map could be developed to allow users to explore projects by price as well as the features that influence property values across different regions in Singapore.

8.0 Ethical Considerations

8.1 Unsupervised Learning

The outlier detection model may draw undue attention to projects with exorbitantly high prices, potentially exacerbating the divide between the rich and the poor in society. This could elicit knee-jerk reactions from policy makers, resulting in poor and ill-informed decisions. Therefore, it is crucial to accompany the model's outputs with additional context, clarifying that these are anomalies rather than the norm in Singapore. An interactive UMAP could also be developed to illustrate the features contributing to each outlier, enabling users to better understand how these cases differ from the population.

8.2 Supervised Learning

The predictions could influence individual buying or selling decisions, which in turn may contribute to market speculation or price inflation. Furthermore, government agencies may utilize the information to inform policy decisions, such as adjusting interest rates to moderate the sales of large private condominiums, given that area is a key predictor of price. To mitigate these concerns, the solution should be accompanied by appropriate disclaimers that communicate the model's limitations, such as its over reliance on area as a feature. Additionally, it would be beneficial to implement a dashboard to enhance transparency and illustrate the features driving its predictions.

9.0 Statement of Work

Table 16: Statement of Work

Project Member	Contribution
Jerome Wong	● Preprocessing data - Distance to train stations ● Unsupervised learning - Outlier detection ● Supervised learning - Failure analysis
Mehmet Balkan	● Preprocessing data - Economic Indicators ● Unsupervised learning - K-Means clustering ● Supervised learning - Feature ablation analysis
Nikoloz Jaghiashvili	● Preprocessing data - Google POI data ● Supervised learning - Trade-Offs ● Supervised learning - Sensitivity analysis

Appendix A: Features Used in Modeling

Features Used in Supervised Learning

Name of Feature	Short Description	Type
age_bin	Age group of the property in bins	Categorical
district	Planning district in Singapore	Categorical
floorRange	Floor level range of the unit	Categorical
marketSegment	Market segment (e.g. Core Central Region)	Categorical
tenure_bin	Lease tenure category of the property	Categorical
typeOfArea	Type of area measurement (e.g. Strata, Land)	Categorical
typeOfSale	Type of sale (e.g. resale, new sale)	Categorical
area	Floor area of the unit in square meters	Numerical
days_since_1st_trans	Days since the first recorded transaction	Numerical
avg_price_level_hospital	Average price level of nearby hospitals	Numerical
avg_price_level_lodging	Average price level of nearby lodgings	Numerical
avg_price_level_police	Average price level of nearby police stations	Numerical
avg_price_level_restaurant	Average price level of nearby restaurants	Numerical
avg_price_level_school	Average price level of nearby schools	Numerical
avg_price_level_shopping_mall	Average price level of nearby shopping malls	Numerical
avg_rating_hospital	Average rating of nearby hospitals	Numerical
avg_rating_lodging	Average rating of nearby lodgings	Numerical
avg_rating_police	Average rating of nearby police stations	Numerical
avg_rating_restaurant	Average rating of nearby restaurants	Numerical
avg_rating_school	Average rating of nearby schools	Numerical
avg_rating_shopping_mall	Average rating of nearby shopping malls	Numerical
cpi_accum	Cumulative Consumer Price Index	Numerical
lrt_nearest_distance_m	Distance to the nearest LRT station (Meters)	Numerical
monthly_marriage_crude_rate	Monthly crude marriage rate	Numerical
monthly_population_growth_rate	Monthly population growth rate	Numerical
monthly_price_index	Monthly price index	Numerical
mrt_nearest_distance_m	Distance to the nearest MRT station (Meters)	Numerical
poi_count_hospital	Number of nearby hospitals	Numerical
poi_count_lodging	Number of nearby lodgings	Numerical
poi_count_police	Number of nearby police stations	Numerical
poi_count_restaurant	Number of nearby restaurants	Numerical
poi_count_school	Number of nearby schools	Numerical
poi_count_shopping_mall	Number of nearby shopping malls	Numerical
SORA	Singapore Overnight Rate Average	Numerical

Features Used in Outlier Detection

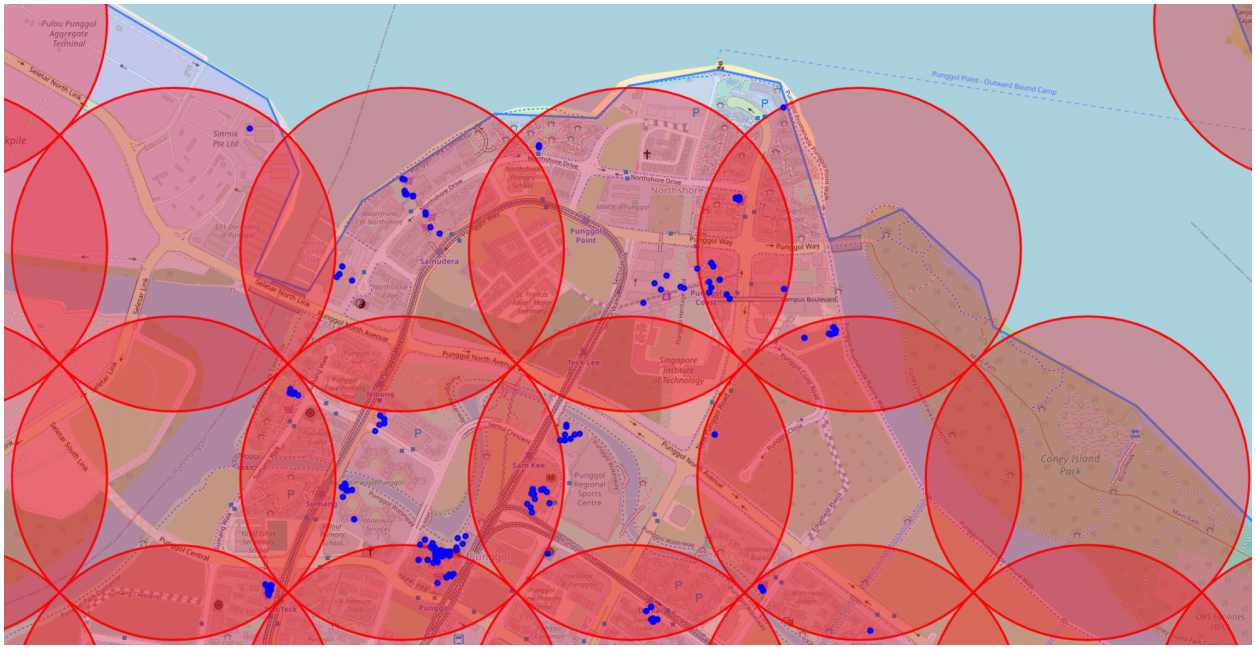
Name of Feature	Short Description	Type
marketSegment	Market segment (e.g. Core Central Region)	Categorical
area	Floor area of the unit in square meters	Numerical
noOfUnits	The number of units in this transaction.	Numerical
target_price	The transacted price nettPrice	Numerical

Features Used in K-Means Clustering

Name of Feature	Short Description	Type
area	Floor area of the unit in square meters	Numerical
lrt_nearest_distance_m	Distance to the nearest LRT station (Meters)	Numerical
mrt_nearest_distance_m	Distance to the nearest MRT station (Meters)	Numerical
poi_count_shopping_mall	Number of nearby shopping malls	Numerical
poi_count_restaurant	Number of nearby restaurants	Numerical
poi_count_police	Number of nearby police stations	Numerical
poi_count_school	Number of nearby schools	Numerical
tenure	Lease tenure	Numerical

Appendix B: Google Place API Querying Radius

Figure 7: Google Place API Querying Radius

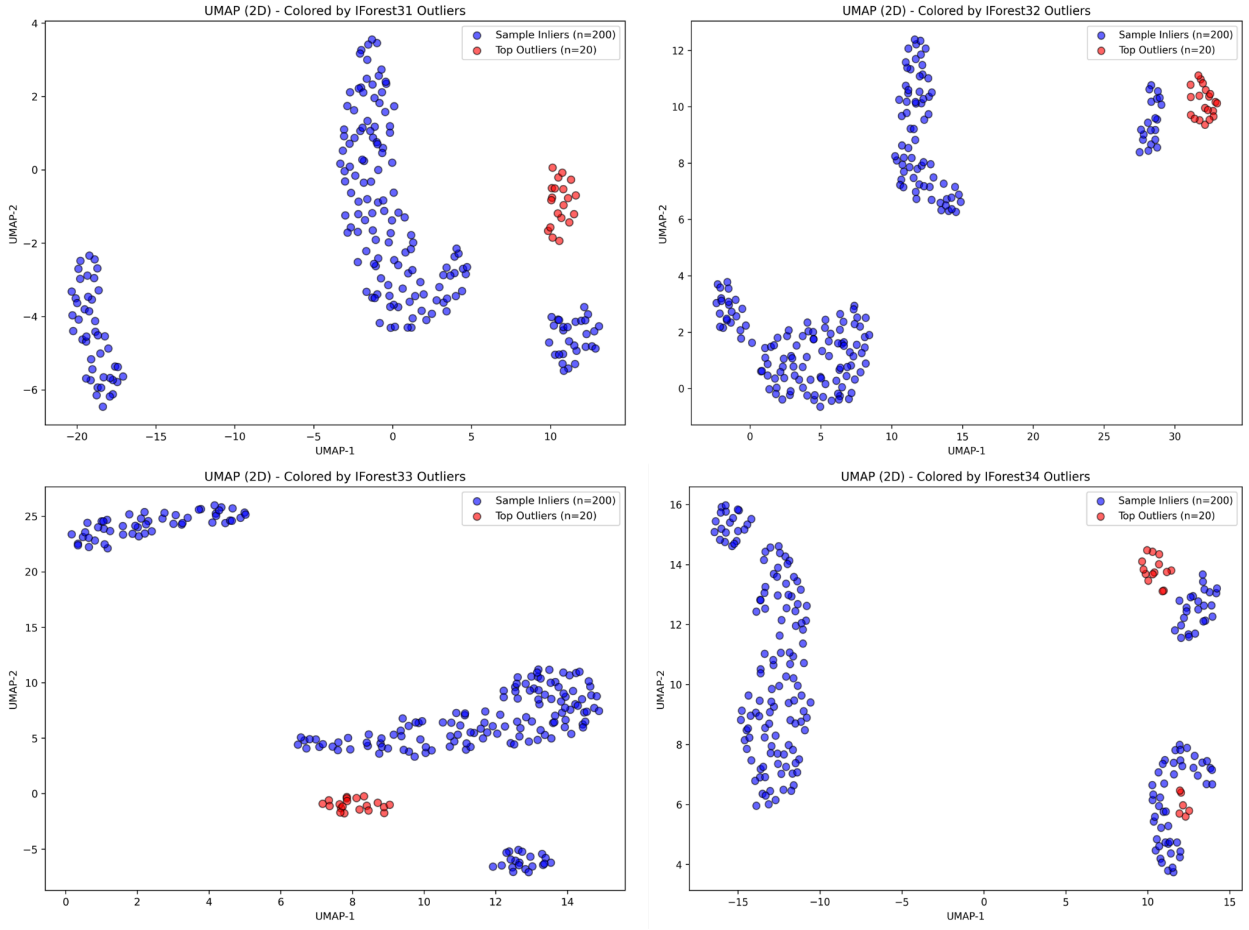


Appendix C: Isolation Forest Sensitivity Analysis Plots

Table 17: Sensitivity Analysis Best Model Settings

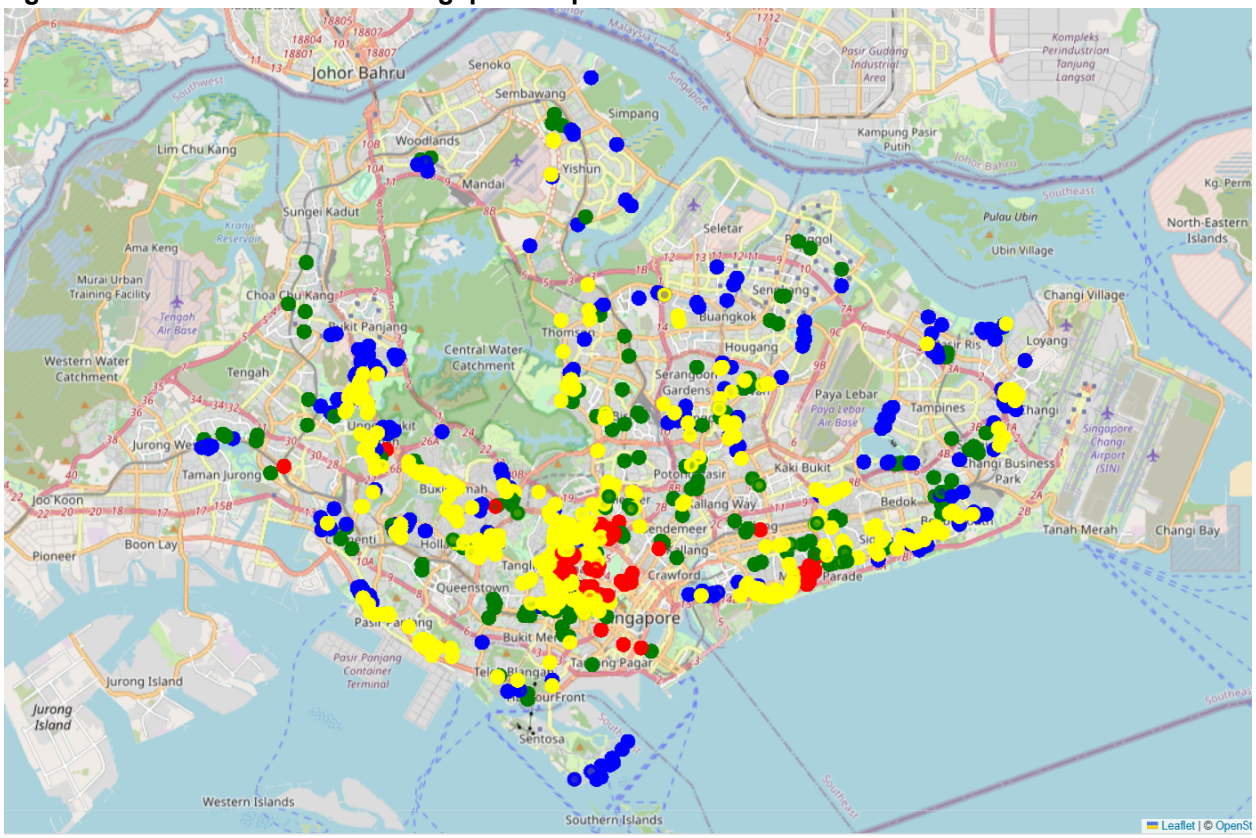
Name	n_estimators	max_features	Best Setting
IForest 31	200	1	Yes
IForest 32	200	2	No
IForest 33	200	3	No
IForest 34	200	4	No

Figure 8: Outlier Detection Sensitivity Analysis



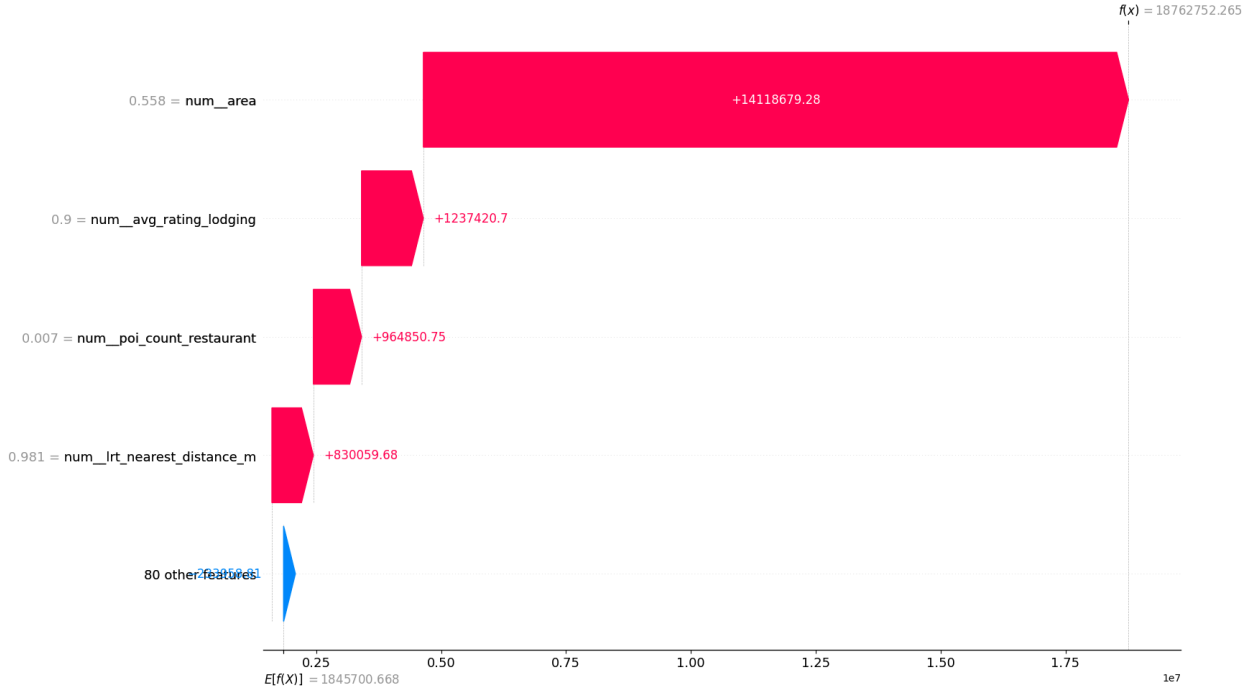
Appendix D: K-Means Cluster Visualization Map

Figure 9: Cluster Visualization Singapore Map



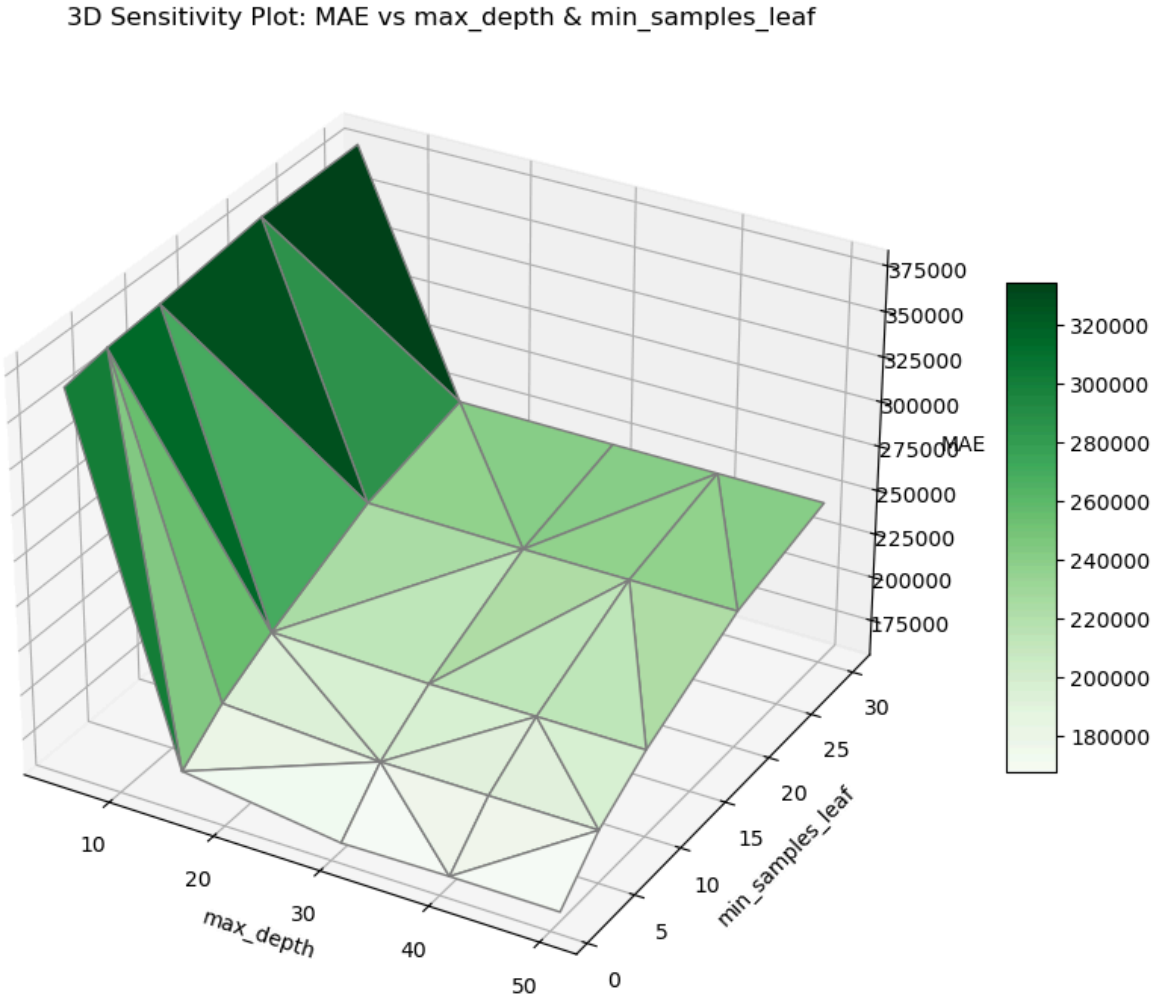
Appendix E: Failure Analysis SHAP Waterfall Plot

Figure 10: Project Turquoise



Appendix F: Sensitivity Analysis

Figure 11: Interaction between max_depth and min_samples_leaf



References

[1] Tay, R. (2024). *Roydontay/SG-property-price-prediction*. GitHub. https://github.com/RoydonTay/SG-Property-Price-Prediction/tree/main

[2] Calis, E. (2023). *ezracalis/Singapore-HDB-Price-Prediction.* GitHub. https://github.com/ezracalis/Singapore-HDB-Price-Prediction

[3] Mallaiyan, G. (2024). *Go7bi/-Singapore-Resale-Flat-Prices-Prediction.* GitHub. https://github.com/Go7bi/-Singapore-Resale-Flat-Prices-Prediction

[4] Leung, Charles Ka Yui and Zhang, Jun and Ma, Wai, The Market Valuation of Interior Design and Developers Strategies: A Simple Theory and Some Evidence (2013). Available at SSRN: https://ssrn.com/abstract=2203647

[5] Mamre, Mari Olsen and Sommervoll, Dag Einar, Coming of Age: Renovation Premiums in Housing Markets (2022). Available at SSRN: https://ssrn.com/abstract=4044171