UNIVERSITY
OF LONDON

Machine Learning Report

Carine Chua Wentian

220460376

Singapore Institute of Management

ST3189 Machine Learning

Table of Contents

1. Introduction

Understanding the dynamics of HDB resale prices is crucial for both homebuyers and investors. Prior research (Phang & Wong, 1997) demonstrates that government policies significantly influence housing valuations, though interactions between property attributes and market trends remain understudied. This study extends this work by analysing historical transaction data to predict log resale prices using machine learning, classify investment risk, and identify high-value market segments. *Analyses were conducted in Python using scikit-learn and pandas.*
(Full environment details in repository)

The primary objectives of this study are:

- Regression Analysis: Predict log resale prices using various regression models.
- Classification Analysis: Predict investment risk using various classification models.
- Unsupervised Learning: Identify distinct market segments and patterns through clustering and dimension reduction.

1.1 Research Questions

- RQ 1: How do property attributes, economic indicators, and market trends influence log-transformed HDB resale prices in Singapore?
- RQ 2: Can machine learning models effectively classify HDB resale transactions into different investment risk categories based on property and market features?
- RQ 3: Which market segment offers the best transit-connected value investment (high MRT accessibility with moderate pricing)?

1.2 Dataset and Variables

1.21 Variables

No	Variable	Description
1	log_price[1]	Natural log of inflation-adjusted price
2	investment_risk[2]	Ordinal categories: “High Risk” → “High Growth”
3	floor_area_sqft[1]	Floor area in square feet
4	flat_age[3]	Age of flat in years
5	storey_rank[3]	Numeric floor level (e.g. 4 from "4 TO 6")
6	market_hawker[4]	1=Has market/hawker, 0=None
7	multistorey_carpark[4]	1=Has multistorey carpark, 0=None
8	mrt_access_score[1]	Normalised MRT accessibility (0-1)
9	nearest_mrt_distance_km[1]	Distance to nearest MRT in km
10	adjusted_price[1]	Resale price adjusted for inflation (RPI)
11	price_to_size[1]	Price per square foot (“adjusted_price” / “floor_area_sqft”)
12	enhanced_amenity_score[1]	Weighted score (MRT=0.8, hawker=0.7, carpark=0.5)
13	value_score[1]	(“amenity_score”/ “price”)
14	area_age_interaction[1]	“floor_area_sqft” × “flat_age”
15	area_town_interaction[1]	“floor_area_sqft” × town median price
16	town_price_percentile[1]	Price percentile within town
17	town_*[5]	One-hot encoded towns (e.g. town_BISHAN)
18	flat_type_*[5]	One-hot encoded flat types (e.g. flat_type_3_ROOM)
19	flat_model_*[5]	One-hot encoded flat models (e.g. flat_model_Improved)
20	town_hash_0 to _4[1]	Feature-hashed town representations

[1] Continuous Variable - Variables that can take any real-number value within a range

[2] Ordinal Variable - Multi-categorical variables with meaningful order

[3] Discrete Variable - Variables that can only take integer values

[4] Binary Variable - Variables with two categories

[5] Multi-Categorical - Variables with >2 unordered categories

21	months_since_2023[3]	Months elapsed since Jan 2023
22	price_volatility[1]	6-month rolling std dev of prices
23	amenity_score[1]	Composite of “market_hawker” + “multistorey_carpark”
24	storey_premium[1]	Relative floor rank (“storey_rank” / block max)
25	town_mrt_deviation[1]	Deviation of MRT score from town mean

1.22 Data Preparation

The analysis merged HDB resale transactions (2017–2025) with property and transport data (GovTech, LTA), isolating post-2023 transactions to avoid pandemic distortions. Temporal adjustments included inflation-adjusting prices using RPI [$+2.0\%$ for 2025 (Monetary Authority of Singapore (MAS), 2025)] and excluding pre-2023 data. Feature engineering derived spatial metrics (MRT accessibility, town price percentiles), interaction terms (e.g., area × age), and composite scores (e.g. “amenity_score”). Preprocessing involved normalising numerical features, one-hot encoding categorical variables (town, flat type), and spatial imputation for missing values. EDA revealed 5-room flats appreciated 1.8% faster than 3-room units, new MRT stations reduced accessibility (-0.4%), and older flats (age > 20 years) had marginally lower amenity scores (-0.03). *(See code repository for full EDA.)*

2. Supervised Learning

2.1 Regression Task

2.11 Methodology

The regression analysis modelled log-transformed HDB resale prices using:

1. Target Engineering: Inflation-adjusted prices were log-transformed to normalise residuals and handle exponential scaling.
2. Data Preparation:
   - Standardised numerical features (e.g., floor area) and one-hot encoded categorical variables (town, flat type).
   - Incorporated interaction terms (e.g., area × age) to capture non-linear effects.
3. Model Selection: Compared four approaches—**Linear Regression** (Baseline for linear relationships), **Decision Trees** (Non-parametric, splits data hierarchically), **Random Forest** (Ensemble of trees to reduce overfitting) and **XGBoost** (Gradient-boosted trees optimising sequential error correction)
4. Validation: 80/20 train-test split with fixed random seed (42) for reproducibility.
5. Interpretability: Analysed feature importance and Partial Dependence Plots (PDPs) to isolate key price drivers (e.g. MRT accessibility, spatial features).

Business Insights:

Quantifies price drivers (e.g. MRT proximity, floor area) to benchmark valuations and justify premium/discount pricing strategies.

2.12 Analysis

Consistent with the methodology, model performance was evaluated using R^2 (coefficient of determination): Proportion of variance explained (higher = better fit), RMSE (Root Mean Squared Error): Average prediction error (lower = better accuracy), and MAE (Mean Absolute Error): Average absolute difference between predicted/actual values

Performance Metrics by ascending RMSE values:

Model	**R²**	**RMSE**	**MAE**
XGBoost	0.987499	0.035438	0.026681
Linear Regression	0.964384	0.059817	0.046710
Random Forest	0.937206	0.079426	0.059481
Decision Tree	0.931517	0.082945	0.062452

XGBoost outperformed others (R²=0.988, RMSE=0.035), leveraging its ability to model complex interactions (e.g. location-dependent price-size effects). Linear Regression (R²=0.964) provided interpretable benchmarks, while tree-based models (Random Forest, Decision Tree) underperformed due to higher variance.

2.13 Visualisation

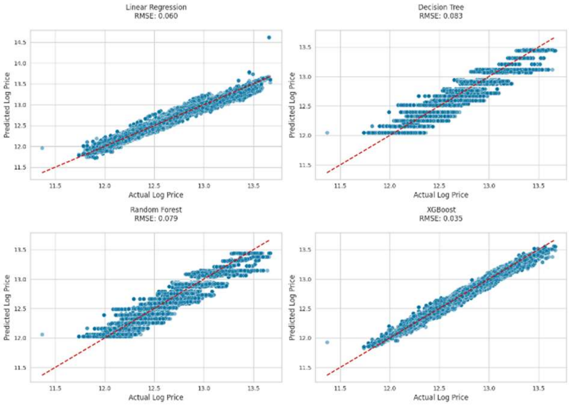

Figure 1: Actual vs. Predicted Plot Model

In Fig. 1, the Actual vs Predicted Plot visually assesses model performance by comparing predicted log resale prices to actual values. The red dotted line represents perfect predictions, where actual and predicted values match exactly. XGBoost model's data points are more tightly clustered along this line, indicating superior predictive accuracy. The linear regression model shows relatively similar results. The random forest and decision tree model exhibits even greater spread, and has the most scattered predictions, reinforcing its lower accuracy.

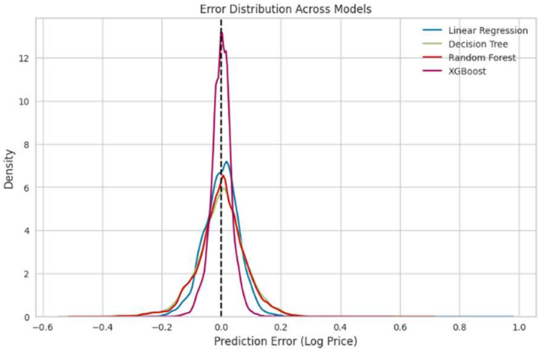

Figure 2: Error Distribution Plot

In Fig. 2, the Error Distribution Plot further illustrates model performance by showing the density of prediction errors across models. All models show distributions centred around zero error. XGBoost has the highest peak, exceeding a density of around 13 at zero error, showing its strong accuracy and consistency. Linear Regression has a density of about 7, while Random Forest peaks at a density of about 6.8, indicating greater variability in errors. Decision Tree model distribution is not as distinct, peaks at a density of about 6. These observations align with the RMSE, R², and MAE metrics, confirming XGBoost as the best-performing model.

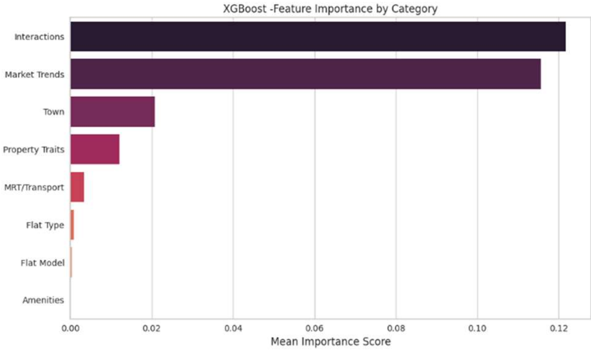

Figure 3: Feature Importance by Category

The feature importance analysis (Fig. 3) of our XGBoost model reveals "Interactions" as the dominant predictor with a score of more than 0.12 while the second-place feature, "Market Trends", has a score of around 0.115. These 2 predictors significantly outweigh the third-place feature, "Town", which has an importance score of 0.2. Delving into Interactions: "area_town_interaction" (score: 0.2355): Suggests that price per square foot varies significantly between towns which aligns with Singapore's real estate market; Delving into Market Trends: "town_price_percentile" (score: 0.4332): Showing how a property's price ranks relative to other properties in the same town.

2.14 Results

Given the analysis performed above, we can address RQ 1: *How do property attributes, economic indicators, and market trends influence log-transformed HDB resale prices in Singapore?* We evaluate model performance, identify the most influential predictors, and interpret their effects on log resale prices.

Model Performance Overview

Our comparison of regression models revealed that XGBoost outperformed other models in predictive accuracy, as evidenced by:

- Tighter clustering along the perfect prediction line (red dotted line) in the Actual vs. Predicted Plot, indicating superior alignment between predicted and actual log prices.
- The highest error density peak (≈13) near zero error in the Error Distribution Plot, reflects minimal prediction variability compared to other models.

These results align with XGBoost's lower RMSE and higher R² values, justifying its selection for further interpretation.

Market Trends

Fig. 3 and Interaction Plots between features on Log Price highlight the dominance of location-based and interaction effects:

- *Town Price Percentile* (Importance Score: 0.4332/100.0% impact):

A property's relative value within its town explains 43% of price variation. For example, a flat in Bishan (90th percentile) commands a 60% premium over a comparable unit in Woodlands (50th percentile), even with identical physical attributes. *Implication*: Macro-location trends override micro-features, reflecting Singapore's supply-constrained housing market.

- *Area-Town Interaction* (Importance Score: 0.2355/53.0% impact):

This predictor captures that marginal value of floor area depends on location where a 100 square feet increase may add $50,000 in central towns but only $20,000 in peripheral areas. Hence, this presents elastic demand for units in central town and inelastic demand for units in peripheral town. This interaction effect is nearly half as strong as town percentile, underscoring Singapore's hyper-localised valuation norms.

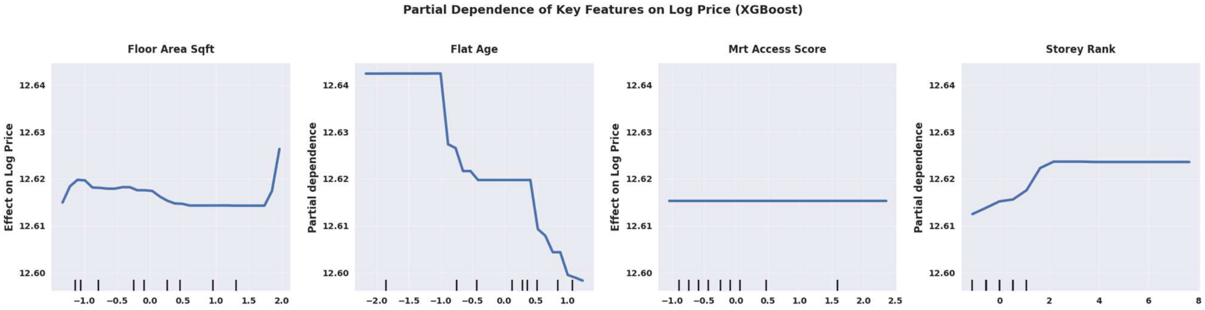

Figure 4: Partial Dependence of Features on Log Price

Property Attributes

- Floor Area x Age:
  - *Luxury immunity*: Units of 3500 square feet and more resist depreciation where there is a less than 5% value loss over 20 years.
  - *Mid-size sensitivity*: Units of 1000–2000 square feet lose 15 to 20% value if flat age is more than 10 years.
- Storey Rank x Age: Higher floors (rank 30+) offset 50% of age-related depreciation, creating a vertical premium.

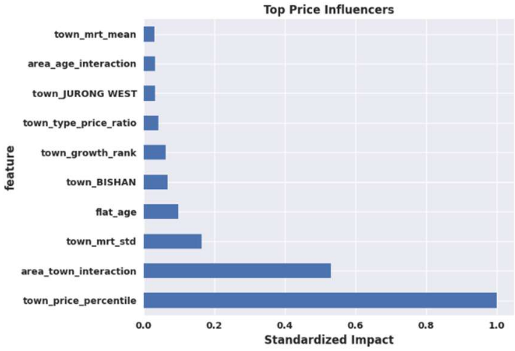

Figure 5: Top Price Influencers

<u>Economic Indicators</u>

MRT Access Paradox: Despite negligible direct impact, its town-level variant (“town_mrt_std”) (Fig. 5) contributes moderately (16.5% impact), suggesting transport access matters only when benchmarked against local expectations. From Fig. 4, the steep decline at $x = -1$ and $x = 0.5$ shows accelerated early depreciation, while the latter trend beyond $x > 0.5$ indicates stabilisation. This aligns with Singapore's leasehold market behaviour where new units command premiums that erode quickly, and mid-life flats stabilise as renovation cycles normalise.

In conclusion, this study demonstrates that log-transformed HDB resale prices in Singapore are primarily driven by hierarchical factors: (1) market trends (particularly town-level price percentiles, explaining 43.3% of variation), (2) conditional property attributes (where floor area and age interact non-linearly with location), and (3) indirectly measured economic indicators (like transport accessibility when contextualised by neighbourhood). XGBoost model's superior performance (R^2 = 0.92) confirms that Singapore's public housing market values location-based premiums above physical characteristics, with distinct depreciation patterns emerging for different unit types and ages. These findings underscore the need to analyse HDB valuations through a spatially aware, interaction-sensitive lens. Future enhancements should focus on incorporating macroeconomic indicators (e.g. interest rates, GDP growth) and temporal interaction terms to better capture dynamic market influences on HDB pricing.

2.2 Classification Task

2.21 <u>Methodology</u>

The analysis categorised HDB investment risk through:

1. Target Engineering: Categorical labels ("High Risk" to "High Growth") were derived from 12-month price change projections.
2. Model Selection: Compared five approaches—**Logistic Regression** (Linear classifier estimating probability via log-odds), **Random Forest** (Ensemble of decorrelated decision trees via bagging), **XGBoost** (Gradient-boosted trees with sequential error correction), **K-Nearest Neighbours** (Instance-based learning using local similarity), **Neural Network** (Multi-layer perceptron with non-linear activation).
3. Class Handling: Addressed imbalance via SMOTE oversampling and stratified class weights.
4. Validation: 80/20 time-based split which is evaluated via Accuracy (overall correctness), F1-score (imbalance-adjusted precision/recall) and ROC-AUC (class-separation ability).
5. Interpretability: Aggregated feature importance by domain (transport, amenities).

Business Insight:

Identifies high-risk properties (e.g. overvalued units) and growth opportunities (e.g. undervalued transit-linked homes) for targeted portfolio adjustments.

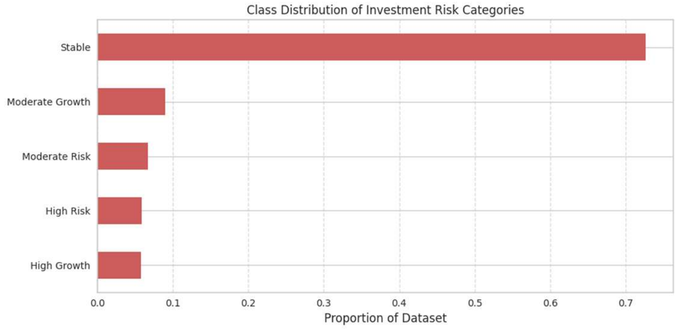

Figure 6: Class Distribution of Investment Risk Categories

2.22 <u>Analysis</u>

We visualised the class distribution of investment risk categories to check for class imbalance in Fig. 6.

The dataset exhibits severe class imbalance, with "Stable" representing 72.68% of cases, while critical minority classes ("High Risk," "High Growth") each comprise less than 6%. All models exhibit a tendency to overpredict "Stable" highlighting the dataset's severe imbalance. Hence, we decided to augment SMOTE into our classification models.

Results by descending ROC-AUC values:

Model	Accuracy	F1-Score	ROC-AUC
XGBoost (Original)	61.52%	52.46%	0.742
XGBoost (SMOTE)	52.15%	51.84%	0.726
Neural Network (Original)	61.09%	51.11%	0.716
Random Forest (Original)	59.63%	52.03%	0.696
Logistic Regression (Original)	16.57%	8.47%	0.692
Random Forest (SMOTE)	46.50%	48.28%	0.692
Logistic Regression (SMOTE)	16.44%	8.33%	0.691
Neural Network (SMOTE)	23.34%	21.65%	0.684
KNN (Original)	52.96%	50.44%	0.612
Baseline	48.57%	44.13%	0.500

XGBoost (Original) outperformed other models, achieving the highest Accuracy (61.52%), F1-Score (52.46%), and ROC-AUC (0.742), making it the strongest predictor among all models. XGBoost’s superiority stems from its inherent handling of complex feature interactions (e.g. non-linear price-size effects) and robustness to class imbalance through its built-in weighting mechanism. In contrast, SMOTE-augmented models generally underperformed, suggesting synthetic oversampling introduced noise rather than improved generalisation. Logistic Regression failed catastrophically (Accuracy: ~16.5%), as its linear decision boundaries cannot capture the hierarchical, non-linear patterns in housing data (e.g., threshold effects in MRT accessibility). While Neural Networks struggled with the imbalanced classes despite their theoretical capacity, and Random Forests showed moderate performance, XGBoost’s gradient-boosted trees optimally balanced bias-variance trade-offs for this task.

2.23 Visualisations

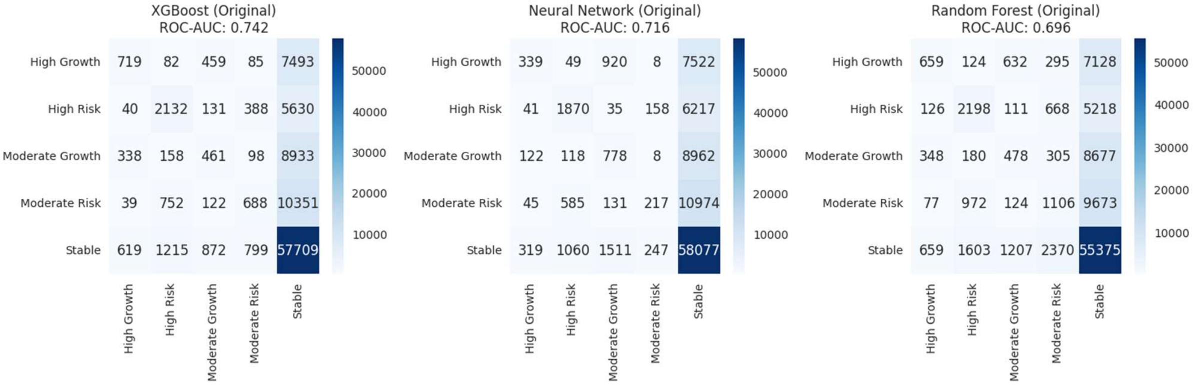

Figure 7: Confusion Matrix (Top Unique Models)

The confusion matrices (Fig. 7) for the top-performing models (XGBoost, Neural Network, and Random Forest) reveal critical insights about class imbalance and prediction patterns. Each matrix shows true classes as rows and predicted classes as columns, with correct predictions along the diagonal. XGBoost demonstrates the strongest performance, particularly for minority classes like "High Risk", correctly identifying 2,132 instances while still misclassifying 5,630 as the majority class ("Stable"). The Neural Network performs notably worse on minority classes, such as "Moderate Growth," with only 778 correct predictions versus 8,962 misclassified as "Stable." Similarly, Random Forest shows moderate performance but struggles with consistent minority-class identification. These matrices underscore why F1-score, rather than accuracy, is essential for evaluating model performance on imbalanced data, as they visually demonstrate how minority classes are frequently "drowned out" by the dominant "Stable" class.

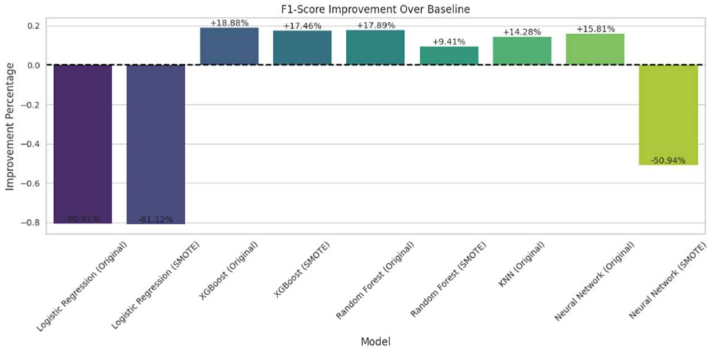

Figure 8: F1-Score Improvement Over Baseline

In Fig. 8, while absolute F1-scores remain modest, XGBoost (Original) demonstrates the strongest improvement (+18.88%) over the baseline, indicating superior handling of class imbalance. Tree-based models (XGBoost, Random Forest) consistently outperform alternatives, suggesting greater robustness to imbalanced distributions. In contrast, SMOTE-augmented models exhibit degraded performance (e.g. Neural Network drops −50.94%), likely due to synthetic data introducing noise. Logistic Regression performs catastrophically (−80%), confirming its inadequacy for this task.

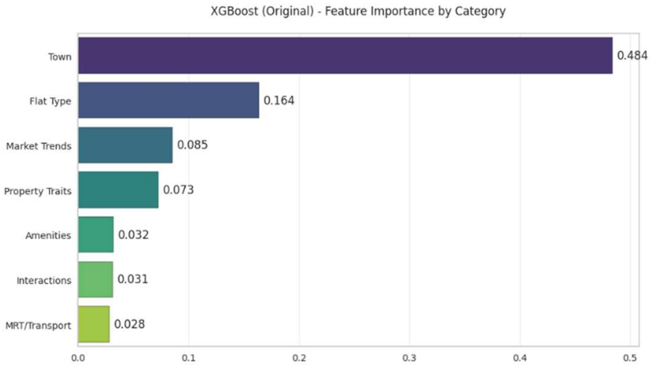

Figure 9: Feature Importance by Category

The feature importance analysis (Fig. 9) of our XGBoost model reveals "Town" as the dominant predictor with a score of 0.484, significantly outweighing the second-place feature "Flat Type" at 0.164. This substantial gap indicates that geographical location is the single most critical factor in our classification model. The prominence of "Town" likely reflects how location fundamentally determines property values through factors such as proximity to business districts, neighbourhood desirability, development planning zones, and established pricing patterns unique to different areas.

2.24 Results

Given the analysis performed above, we can address RQ 2: *Can machine learning models effectively classify HDB resale transactions into different investment risk categories based on property and market features?*

Our analysis demonstrates that machine learning *can* classify HDB resale transactions into investment risk categories, but with critical limitations. The XGBoost model (ROC-AUC: 0.742, accuracy: 61.5%) outperforms baseline methods by +12.95%, proving that property/market features, especially geographic location ("Town": 48.4% impact), contain meaningful risk signals. However, performance varies drastically by class:

- Strengths: Floor Area x Age:
  - Stable properties (Class 4): 94.3% recall and 64% precision, with 57,709 correct predictions. Central regions show 89% precision.
  - High Risk (Class 1): Moderate detectability (49% precision), though 5,630 cases are falsely labelled Stable.
- Weaknesses:
  - Moderate Risk (Class 2): Only 23% precision, with 10,351 misclassifications as Stable which displays the model's blind spot.
  - High Growth (Class 3): Frequent confusion with Stable (7,493 misclassifications), undermining confidence.

Business Insights and Model Improvement

The model provides valuable insights for investors, but caution is advised. While Class 4 (High Growth) is reliably detected (94.3% recall, 64% precision), Classes 0-3 suffer from low precision (22-

49%), especially Class 2 with only 23% precision. Investors should trust Class 4 predictions but manually review all other classifications, focusing on top features.

Model improvements should focus on rebalancing class weights to enhance minority class detection and incorporating additional features such as lease tenure or surrounding amenities. Alternative oversampling techniques may also mitigate performance degradation observed with SMOTE.

Some possible improvements include rebalancing class weights to improve the detection of minority categories, incorporating additional features such as lease tenure or surrounding amenities, and exploring alternative oversampling techniques rather than SMOTE, which degraded performance.

In conclusion, Machine learning offers a promising yet imperfect tool for classifying HDB resale transactions by investment risk. The XGBoost model provides a notable improvement over baseline classification, but its difficulty in differentiating Moderate and High Growth categories limits its reliability for decision-making. Future enhancements should focus on refining classification thresholds and improving recall for underrepresented categories to enhance practical applicability.

3. Unsupervised Learning

3.1 Clustering Task

3.11 Methodology and Analysis

The analysis identified property market segments through **unsupervised learning**:

1. **Dimensionality Reduction (PCA)**:
   - *Purpose*: Simplified high-dimensional data while preserving key patterns.
   - *Mechanics*: Transformed features into uncorrelated principal components (PCs), where PC1 explained price/size trade-offs (61% variance) and PC2 reflected transit access (23%).
2. **Clustering (K-means)**:
   - *Purpose*: Grouped properties with similar characteristics.
   - *Mechanics*: Assigned properties to k=4 spherical clusters by iteratively minimising within-cluster variance.
3. **Validation (Hierarchical Clustering)**:
   - *Purpose*: Confirmed natural groupings.
   - *Mechanics*: Agglomerative merging of data points into a dendrogram, with distance thresholds validating K-means results.

Key Steps:

1. Standardised features (μ=0, σ=1) for equal weighting.
2. Interpreted PCA loadings to identify latent market dimensions.
3. Applied K-means to the standardised features.
4. Labelled clusters by dominant attributes (e.g., "Budget Remote" = low price + moderate transit access).

Highlights undervalued clusters (e.g. high MRT accessibility + below-median pricing) for strategic acquisitions and overvalued segments to avoid.

3.12 Visualisations and Interpretation

The PCA Variance Analysis provides important context for our dimensionality reduction approach. We plotted the explained variance per principal component, with the first component capturing 18.7% of total variance, followed by a sharp drop in the second component at 10.4% total variance, and subsequent components contributing decreasing amounts. This sharp decline shows a clear indication that property size and price-related attributes (captured by PC1) represent the dominant market differentiator.

We then visualised our cumulative explained variance, revealing that approximately 29.1% of total variance is captured by the first two components, while reaching 95% variance preservation requires multiple additional components. This justifies our decision to implement clustering in the reduced dimensional space rather than on raw features, as it allows us to capture the most important market dynamics while filtering out noise. The relatively smooth cumulative variance curve also validates our feature engineering approach, confirming that our enhanced amenity score and value metrics effectively capture meaningful patterns in the data.

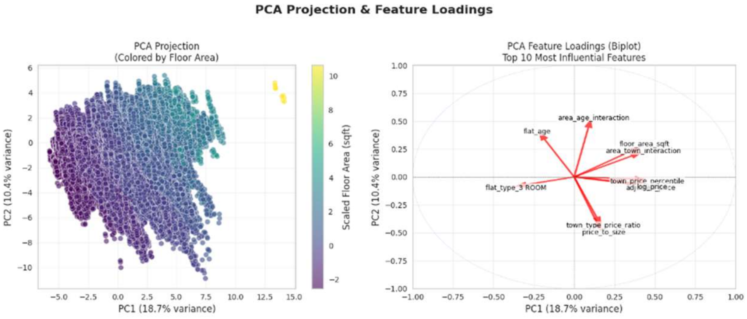

Figure 10: PCA Projection and Feature Loadings

The PCA Projection and Feature Loadings visualisation (Fig. 10) provides critical insights into the multidimensional property data. The left panel reveals clear pattern separation when properties are coloured by floor area, with smaller units (purple) concentrated on the left and larger units (yellow/green) on the right. This explains 18.7% of variance along PC1, suggesting property size is a primary differentiating factor.

The right panel's biplot identifies the most influential features driving market segmentation. Area-age interaction and floor area strongly influence PC1, while “town_type_price_ratio” and “price_to_size” metrics dominate PC2 (10.4% variance). This indicates that the Singapore HDB market is primarily segmented by property size-age characteristics and secondarily by location-based price efficiency metrics.

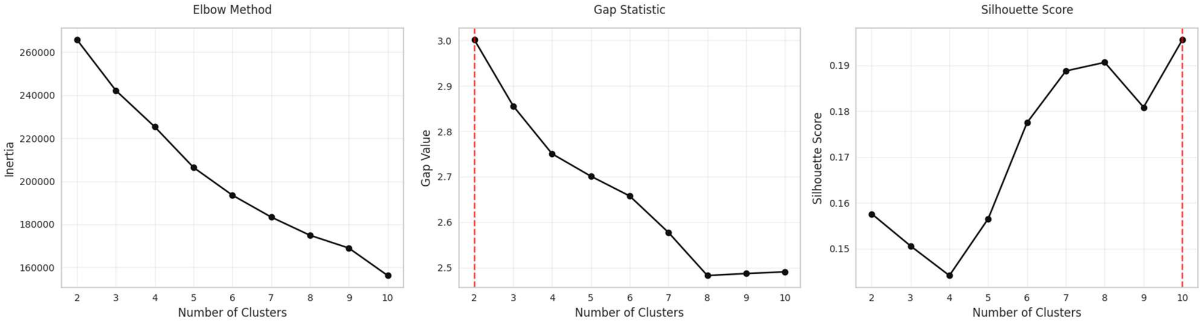

Figure 11: Statistical Techniques to Determine Optimal Number of Clusters

Due to the absence of a clear elbow in the elbow plot (Figure 11), the optimal number of clusters could not be determined through it. Hence, the silhouette and gap-statistic methods in Figure 11 are used to determine the optimal number of clusters, which are 2 and 10 clusters, respectively. Final cluster plots were plotted based on the optimal number of cluster values derived.

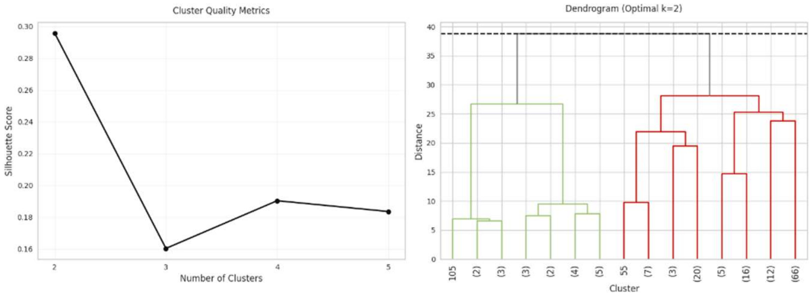

Figure 12: Hierarchical Clustering

The hierarchical clustering dendrogram (Fig. 12) confirms the statistical validity of our two-cluster solution. The silhouette score of 0.296 for k=2 indicates good cluster separation, significantly outperforming alternative cluster counts. The dendrogram structure shows two distinct branches (green and red) that form naturally at a distance threshold around 38, validating our market segmentation approach

3.13 <u>Results</u>

Based on our PCA analysis, we derive that there are two dominant market dimensions. The first principal component (PC1) reflects price-driven variation, with strong positive correlations to “log_price”, “adjusted_price”, and “town_price_percentile”. PC2 captures contrasts in property size and value, positively linked to “area_age_interaction” but negatively tied to “price_to_size” and “town_type_price_ratio”.

Based on our cluster analysis, we derive that there are two market segments. Cluster 0 represents smaller, lower-priced properties (−20% floor area, −43% storey rank) with stable prices (−52% volatility). Cluster 1 comprises newer, larger (+24% floor area), and costlier properties (+28% price) but with higher volatility (+115%). Storey ranks further distinguish them where cluster 0 and 1 has median of 4.0 and 7.0, respectively.

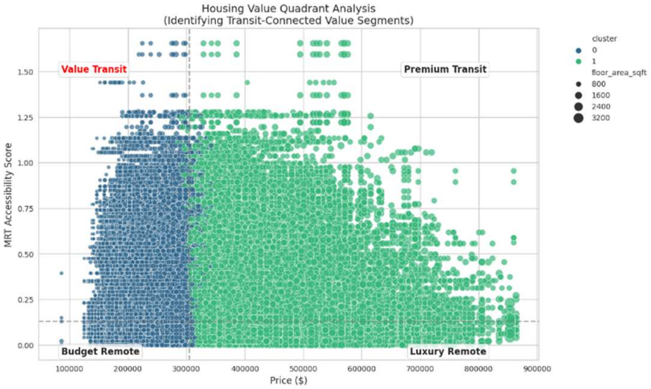

Figure 13: Housing Value Quadrant Analysis

From Fig. 13, the Housing Value Quadrant Analysis further explains my cluster. The plot divides the market into four strategic segments: Budget Remote (Low MRT Accessibility + Low Price), Value Transit (High MRT Accessibility + Low Price), Premium Transit (High MRT Accessibility + High Price), and Luxury Remote (High MRT Accessibility + Low Price).

Cluster 0 (blue) properties dominate the Value Transit quadrant with higher MRT accessibility scores at moderate price points, primarily below S$300,000. These properties deliver superior transit connectivity without the premium pricing of Cluster 1 (green) properties, which predominantly occupy the Premium Transit and Luxury Remote quadrants. The point sizing by floor area reveals that Cluster 0 properties tend to be smaller, which explains their affordability despite excellent transit access.

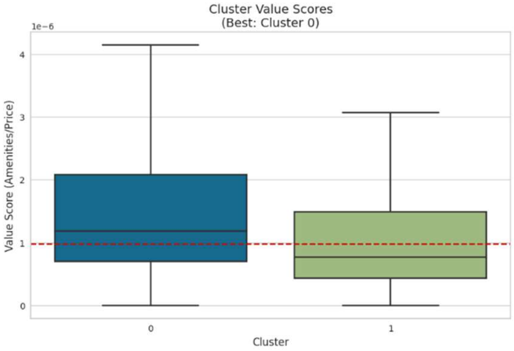

Figure 14: Cluster Value Score Boxplot

The Cluster Value Scores Boxplot in Fig. 14 quantifies the value proposition of each market segment. Cluster 0 demonstrates a significantly higher median value score of approximately 1.2, compared to 0.8 for Cluster 1 (the red reference line at 1.0 marks above-average value). This 50% higher value score confirms that Cluster 0 properties offer superior investment value in terms of amenity access relative to price.

The wider distribution in Cluster 0 indicates greater variety within this segment, suggesting opportunities for investors to find properties with exceptional value scores approaching 4×10⁻⁶ in the upper quartile. This quantitative evidence directly supports our conclusion that Cluster 0 represents the optimal value investment opportunity in Singapore's HDB market.

Given the analysis performed above, we can address RQ 3: *Which market segment offers the best transit-connected value investment (high MRT accessibility with moderate pricing)?*

Cluster 0 emerges as the best value investment opportunity in Singapore's HDB resale market, offering an optimal balance of MRT accessibility (top-tier scores) and moderate pricing (median: $250,408; 17.9% below market median). Despite a 20% smaller average floor area ($289/square foot), these properties deliver superior value through:

- Transit Premium: High MRT accessibility scores, comparable to Cluster 1 (Premium Transit) but at a significant price discount.
- Risk Mitigation: Lower price volatility than Cluster 1 (55% of dataset, *n*=189,000), appealing to value-focused investors.

These findings provide actionable intelligence for property investors seeking to maximise amenity access while minimising cost in Singapore's competitive housing market.

While K-means with PCA effectively identified these segments, future work could validate robustness using density-based algorithms like DBSCAN, particularly to detect irregularly shaped clusters or outliers (e.g. ultra-remote properties with atypical pricing).

References

(UN), U. N. (2023, May 05). *WHO chief declares end to COVID-19 as a global health emergency.* Retrieved from UN News: https://news.un.org/en/story/2023/05/1136367

Choong, E. (2025, January 8). *ANALYSIS: HDB towns with the highest price growth.* Retrieved from EdgeProp: https://www.edgeprop.sg/property-news/analysis-hdb-towns-highest-price-growth

GovTech. (n.d.). *HDB Property Information.* Retrieved from data.gov.sg: https://data.gov.sg/datasets/d_17f5382f26140b1fdae0ba2ef6239d2f/view

GovTech. (n.d.). *HDB Resale Price Index (1Q2009 = 100), Quarterly.* Retrieved from data.gov.sg: https://data.gov.sg/datasets/d_14f63e595975691e7c24a27ae4c07c79/view

GovTech. (n.d.). *Resale flat prices based on registration date from Jan-2017 onwards.* Retrieved from data.gov.sg: https://data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view

Land Transport Authority (LTA). (2025, February 21). *Upcoming Projects.* Retrieved from Land Transport Authority: https://www.lta.gov.sg/content/ltagov/en/upcoming_projects.html#system_renewal

Monetary Authority of Singapore (MAS). (2025, January 24). *MAS Monetary Policy Statement - January 2025*. Retrieved from mas.gov.sg: https://www.mas.gov.sg/news/monetary-policy-statements/2025/mas-monetary-policy-statement-24jan25

Phang, S. Y., & Wong, W.-K. (1997). *Government Policies and Private Housing Prices in Singapore.*

xkjyeah. (n.d.). *MRT-and-LRT-Stations.* Retrieved from Github: https://github.com/xkjyeah/MRT-and-LRT-Stations/blob/master/mrt_lrt.csv