CS5228 Final Report —— Team 2

Predicting HDB Resale Prices with Integrated Temporal, Spatial, and Economic Features

1st Zhao Chujun
Master of Computing
School of Computing
National University of Singapore, Singapore
e1553369@u.nus.edu

2nd Cui Jianghao
Master of Computing
School of Computing
National University of Singapore, Singapore
e1553626@u.nus.edu

3rd Yang Kaiming
Master of Computing
School of Computing
National University of Singapore, Singapore
e1326312@u.nus.edu

4th Li Zimo
Master of Computing
School of Computing
National University of Singapore, Singapore
e1539199@u.nus.edu

***Abstract*—This project focuses on building a robust regression framework to predict the resale prices of HDB flats in Singapore.**

Starting from comprehensive exploratory data analysis, we examined both categorical and numerical attributes to understand their relationships with housing value. Through data cleaning, feature engineering, and integration of auxiliary datasets (including geographic coordinates and macroeconomic indicators such as COE, CPI, SORA, STI, and GDP growth), we enhanced the model's ability to capture spatial, temporal, and economic effects.

Three gradient boosting models—XGBoost, LightGBM, and CatBoost—were trained and compared using cross-validation. Among them, CatBoost achieved the best accuracy, while LightGBM demonstrated a strong trade-off between speed and performance. Feature importance analysis revealed that both intrinsic flat attributes and external economic indicators significantly influence resale prices.

The project demonstrates that integrating multi-dimensional contextual data can substantially improve housing price predictions and provide interpretable insights into Singapore's real estate market.

***Index Terms*—HDB resale prices, feature engineering, gradient boosting, CatBoost, LightGBM, macroeconomic indicators, Singapore housing market**

I. Motivation and Objectives

The Housing & Development Board (HDB) flat resale market is a cornerstone of Singapore's real estate landscape. For both buyers and sellers, a comprehensive understanding of the factors that determine a flat's market value is essential for making well-informed financial decisions. While attributes like floor area, flat type, and proximity to amenities such as MRT stations and malls are commonly accepted as price influencers, their quantitative impact is not always clear.

The primary objective of this project is to address this by developing a robust regression model to predict the resale price of HDB flats based on their given properties. Beyond simply creating a predictive too, this project will also seek to identify and quantify the importance of various flat attributes in determining resale value. A key component of this work involves the evaluation and comparison of different regression techniques to ascertain the most effective approach for this task. The investigation will conclude with a detailed error analysis to understand the model's predictive strengths and weaknesses, along with a discussion on the limitations of the current analysis and potential extensions for future work.

II. Exploratory Data Analysis (EDA) and Preprocessing

The dataset provides comprehensive details regarding HDB resale transactions in Singapore, encompassing location-specific information, property characteristics, and the final resale price. Our objective is to predict the "RESALE_PRICE" by leveraging this information, which consists of both categorical and numerical attribute types.

A. Data Overview and Initial Insights

1) Category Attributes: To intuitively grasp the data's characteristics, we begin by exploring its distribution through visual representations. Pie charts have been generated to analyze the key categorical attributes:"TOWN", "FLAT_TYPE", and"FLAT_MODEL".

An analysis of these visualizations provides valuable insights into the dataset's composition. We observe a relatively balanced distribution for the location-based attribute "TOWN". However, the attributes that describe the physical properties of the flats,"FLAT_TYPE" and "FLAT_MODEL", demonstrate

significant imbalances. For example, in the “FLAT_MODEL” category, “model a” and “improved” flats make up a substantial majority of the data. Similarly, “3-room”, “4-room”, and “5-room” flats are the most dominant categories within “FLAT_TYPE”. This pronounced imbalance is an important finding, as it might lead to larger prediction discrepancies for the less frequently occurring flat types and models during the forecasting phase.

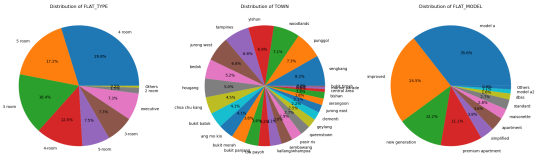

Fig. 1. Visualization of Categorical Data Distribution.

Then, to better understand the factors influencing property value, we have chosen to analyze the price per square meter (“PRICE_PER_SQM”), derived from “RESALE_PRICE” and “FLOOR_AREA_SQM”. By normalizing the price by floor area, “PRICE_PER_SQM” provides a more standardized metric that allows for a fairer comparison of value across different property sizes and types. This helps to isolate the influence of other categorical attributes. We analyzed the “PRICE_PER_SQM” distribution across various categorical attributes and visualized the results using boxplots.

Based on the fig. 2, it appears that the influence of different towns on “PRICE_PER_SQM” is relatively small. The overall distributions across most towns show considerable overlap and do not suggest that location is the most dominant factor. The fig. 3 boxplot, when viewed in conjunction with the distribution pie chart, offers a key insight. The majority of flat models, which represent the bulk of the dataset, exhibit a relatively narrow and similar range of “PRICE_PER_SQM”. However, the rarer models, such as ‘dbss’, ‘terrace’, ‘type s1’, and ‘type s2’, consistently show a significantly higher median “PRICE_PER_SQM”. This indicates that while standard flat models have comparable unit prices, scarcity and unique designs command a substantial price premium, despite their small representation in the overall data.

Interestingly, the fig. 4 reveals a significant finding. We observe seemingly similar categories, such as ’2 room’ and ’2-room’, which demonstrate distinct “PRICE_PER_SQM” distributions. This dissimilarity might arise from differing data sources or classifications over time , potentially representing different standards or sub-categories of properties that are grouped under a similar name. The fig. 4 also shows that a simple correlation between the number of rooms and price per square meter does not hold. While flat types like ’executive’ and ’multi generation’ have a high total resale price, their “PRICE_PER_SQM” is not the highest. Conversely, smaller units such as ’1 room’ and ’2 room’ flats show a relatively high median “PRICE_PER_SQM”. This could be attributed to their typical locations in mature estates with high amenity density or their appeal in the rental market.

2) Numeric Attributes: Alongside the examination of categorical features, we have analyzed the distributions of key numerical attributes: ‘FLOOR_AREA_SQM’, ‘LEASE_COMMENCE_DATE’, and ‘MONTH’. As shown in the ”Distribution of FLOOR_AREA_SQM” in the fig. 5, the ‘FLOOR_AREA_SQM’ attribute exhibits several distinct, sharp peaks rather than a smooth curve. This suggests standardized flat sizes, with high concentrations around specific areas (e.g., $\tilde{7}0$, $\tilde{9}5$, $\tilde{1}10$ sqm), which likely correspond directly to the main ‘FLAT_TYPE’ categories such as 3-room, 4-room, and 5-room. The ”Distribution of LEASE_COMMENCE_DATE” plot in the fig. 5 also highlights non-continuous data. It is clear that HDB construction occurred in distinct clusters or phases, with major building peaks visible around the mid-1980s, the late-1990s, and the mid-2010s. The ”Distribution of MONTH” plot in the fig. 5 shows that the density of transaction records after 2021 is higher than before 2021. This may be due to the more active home buying market after 2021, or it may simply be due to the uneven distribution of data sources over time.

The 2D Kernel Density Estimation (KDE) plots reveal the correlations between these numerical attributes and the ‘PRICE_PER_SQM’. A clear positive correlation is evident for both ‘LEASE_COMMENCE_DATE’ (0.403) and ‘MONTH’ (0.508), which aligns with expectations. Newer lease commencement years and more recent transaction months tend to correlate with higher unit prices. The ‘LEASE_COMMENCE_DATE’ plot in the fig. 6 shows density hot-spots shifting from 4000-6000 SGD for older flats (1980s-1990s) to over 6000 SGD for newer flats (post-2010). The ‘MONTH’ plot shows an even stronger upward-right trend, indicating a significant rise in unit prices across the observed period (2017-2025), likely reflecting market inflation.

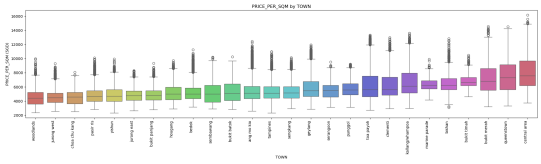

Fig. 2. Boxplot of price per SQM by different town.

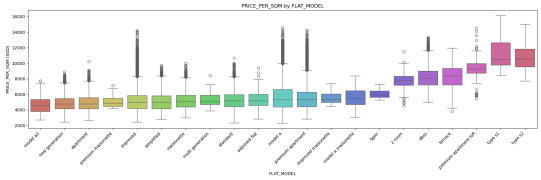

Fig. 3. Boxplot of price per SQM by different model.

3) Assessing Data Quality: Training set and testing set show no missing values. This solid data integrity streamlines preprocessing, allowing a smooth transition into exploratory data analysis and model development.

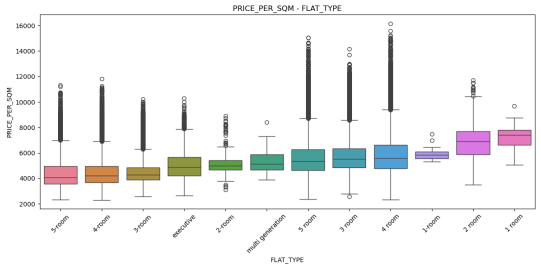

Fig. 4. Boxplot of price per SQM by different type

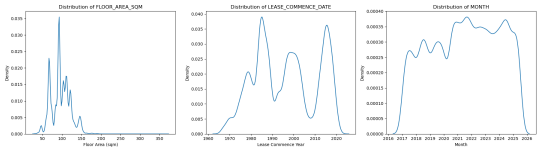

Fig. 5. Distribution of Numerical Features.

Observations from the previously generated boxplots and KDE plots reveal the presence of numerous outliers in the dataset, especially for the 'PRICE_PER_SQM' attribute. These extreme values can significantly impact data analysis and modeling. Such outliers can bias model parameters, leading to overreactions and affecting the model's adaptability to new, unseen data.

To address this, we employed a standard outlier removal approach. We implemented a function that filters the dataset based on a standard deviation threshold for key numerical columns. Specifically, this method was applied to 'FLOOR_AREA_SQM' and 'PRICE_PER_SQM'. Our threshold was set at 3; any data point in these columns falling outside of 3 standard deviations from the mean was considered an outlier and subsequently removed from the training set.

4) Spatial Distribution: The dataset includes several geographic attributes, such as 'TOWN', 'BLOCK', and 'STREET', which are crucial for determining property value. In our preliminary analysis, we observed that while a general price trend exists from outlying to central towns, the 'TOWN' attribute is likely too coarse to capture the significant price variations that occur within these large areas.

While attributes like 'BLOCK' and 'STREET' offer more granular detail, using them directly can be challenging. Therefore, to model the impact of fine-grained geographic location, we have engineered a new feature. We noted that the 'STREET' attribute provides a good balance between granularity and data density, with most streets averaging 100-200 transaction records. Based on this, we created a new attribute, 'AVG_STREET_PPS', which represents the average 'PRICE_PER_SQM' for each street. This approach serves as a computationally efficient proxy that simulates a K-Nearest Neighbors (KNN) model with a large K (K $\approx$ 100 - 200), using the street as a natural geographic cluster. This new feature aims to capture localized price dynamics more effectively than a broad 'TOWN' label.

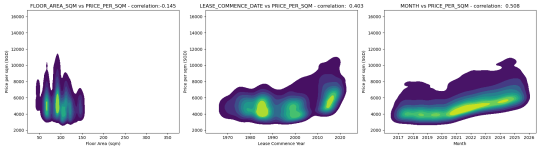

Fig. 6. KDEs with unit price.

5) Temporal Variables: The primary time-related feature in our dataset is 'MONTH', capturing transactions from 2017 to 2025. To gain a detailed understanding of temporal price changes, we first meticulously plotted the distribution of 'PRICE_PER_SQM' across various years. As shown in the fig. 7, this analysis revealed substantial differences in rental price distributions. From 2017 to 2020, the distrubution curves shrift little. However, there is a clear and consistent shift of the entire density curve to the right from 2021 to 2025. This indicates a strong and steady increase in the overall 'PRICE_PER_SQM' over this period.

Furthermore, we delved into a more granular monthly breakdown, as shown in the fig. 8. This visualization clearly confirms the price inflation trend, with both the mean (blue line) and median (green line) unit prices showing a strong upward trajectory, especially accelerating from 2020 onwards. This plot also reveals considerable fluctuations and a general upward trend in the standard deviation (Std, red line), suggesting that price volatility and the spread of property values have also increased over time. These findings underscore that time is a critical predictive factor, and models must account for this significant temporal variance.

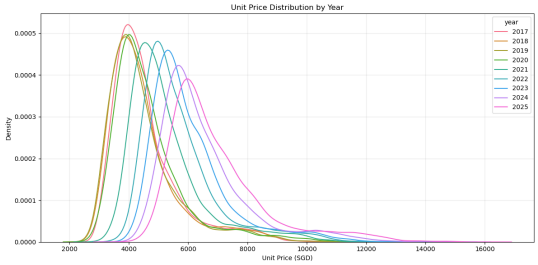

Fig. 7. Unit Price Distribution by Year.

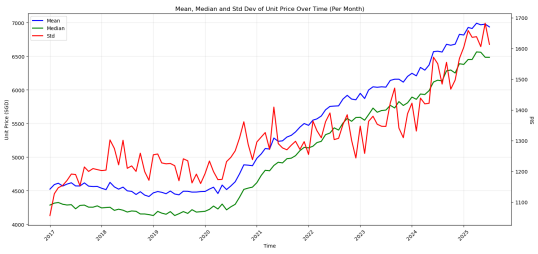

Fig. 8. Mean, Median and Std Dev of Unit Price Over Time.

B. Data Cleaning and Feature Engineering

1) Data Cleaning: Before proceeding with feature transformation, a series of data cleaning operations were performed to ensure data consistency and quality.

First, we removed the 'ECO_CATEGORY' attribute. During the initial analysis, it was discovered that all values for this attribute were "uncategorized", meaning it carried no useful information and was therefore dropped from both the training and test sets.

Additionally, the 'FLOOR_RANGE' attribute (e.g., "01 to 03") is ordinal by nature, making it difficult for models to use directly. We transformed this range into a numerical feature, 'FLOOR_AVG'. By extracting the upper and lower bounds of the range, we calculated the mean value (e.g., "01 to 03" becomes 2.0), converting it into a continuous variable that is easier for the model to process.

2) Transform of Category Attributes: Building on the cleaned data, we transformed the categorical attributes so they could be utilized by the models. For 'FLAT_TYPE' and 'STREET', both high-cardinality categorical attributes, we adopted a Target Encoding strategy. We avoided One-hot encoding as it would have produced excessively high dimensionality.

1. FLAT_TYPE_ENCODED: We calculated the mean 'PRICE_PER_SQM' corresponding to each 'FLAT_TYPE' in the training data. These mean values were then mapped back to the dataset, creating the new 'FLAT_TYPE_encoded' feature.

2. AVG_STREET_PPS: We applied the same target encoding methodology to the 'STREET' attribute. We calculated the average 'PRICE_PER_SQM' for each 'STREET' and added this as a new feature, 'AVG_STREET_PPS'. As mentioned previously, this feature serves as an effective and computationally efficient proxy for the impact of geographic location on price.

3. SPECIAL_MODEL: In our exploratory data analysis (EDA), we observed a significant price distribution difference between the majority of common models (like 'model a') and the few rare, premium models (like 'terrace', 'dbss', etc.) within the 'FLAT_MODEL' attribute. To capture this non-linear relationship, we engineered a new binary feature called 'SPECIAL_MODEL'. We defined a list containing all rare and premium models; if a flat's 'FLAT_MODEL' was in this list, the 'SPECIAL_MODEL' feature was set to 1, otherwise, it was set to 0. This helps the model differentiate between standard units and those that command a special price premium.

C. Auxiliary Data Integration

1) Geographic Data: The initial training dataset contained address information, such as town and block number, but lacked precise geographic coordinates. To address this, our first step was a data integration task. We utilized an auxiliary dataset, `sg-hdb-block-details.csv`, which contained the latitude and longitude for HDB blocks.

Before merging, a critical data preprocessing step was performed to ensure a high matching success rate. The joining keys "TOWN" and "BLOCK" in both the primary and auxiliary datasets were normalized. This involved converting all text to lowercase and removing any leading or trailing whitespace, creating a standardized key for a robust join operation. This process was validated and achieved a 100% success rate, providing a foundational geographic layer for all subsequent feature engineering.

With the latitude and longitude for each property established, we proceeded to the feature engineering phase. We selected three types of amenities critical to the Singaporean residential landscape: MRT stations, hawker centres, and primary schools.

For each property, we calculated two distinct types of features for each amenity category:

- Distance to the Nearest Amenity: This feature captures the direct convenience of access. It was computed by calculating the geodesic distance between a property and every amenity of a given type, then selecting the minimum value. The geodesic distance provides an accurate, "as-the-crow-flies" measure over the Earth's surface.
- Count of Amenities within a Defined Radius: This feature measures the density and availability of options in the vicinity of a property. A higher count may suggest a more vibrant and well-served neighborhood.

D. Economic and Temporal Feature Construction and Analysis

1) Data Sources and Integration: In addition to the physical and location attributes of each flat, the broader economic environment also has a notable impact on housing prices in Singapore. Economic conditions such as inflation, interest rates, and overall growth can influence people's purchasing power, investment decisions, and consequently both resale and rental prices. To capture these external effects more accurately, we introduced several macroeconomic and temporal features into our analysis.

For the **temporal features**, we relied mainly on the HDB resale dataset itself. From each transaction record, we extracted the year, month, and quarter information, and used the `lease_commence_date` to calculate building age and remaining lease years. These attributes help capture long-term market trends and the depreciation effects of older flats.

For the **economic features**, we collected multiple public datasets that reflect different aspects of Singapore's economic landscape:

- **COE (Certificate of Entitlement):** Monthly certificate prices and quotas, representing consumer confidence and economic activity.
- **CPI (Consumer Price Index):** Monthly inflation indicator published by the Singapore Department of Statistics.
- **SORA (Singapore Overnight Rate Average):** Daily market interest rate from the Monetary Authority of Singapore, averaged by month.
- **STI (Straits Times Index):** Daily stock market index from Yahoo Finance, aggregated into monthly mean and last-day close.
- **GDP Growth Rate:** Quarterly real GDP growth rates released by official statistics.

Since these data sources use different time granularities (for example, GDP data is quarterly while others are monthly), we standardized all of them to a common **monthly format**. Dates were converted to the first day of each month (e.g., "2020-10-01"), and each dataset was merged with the main HDB table using the `month` field via a left join. This alignment ensures that every economic indicator corresponds correctly to the month of each transaction.

By combining these temporal and macroeconomic datasets, our model can account for both intrinsic housing attributes and broader external conditions, making its predictions more robust and better aligned with real market behavior.

2) Feature Construction and Transformation: After aligning all datasets, we created two main groups of features: **temporal features** from the main HDB dataset and **economic features** from external data sources. These features enable the model to capture both time-related patterns and broader economic influences on housing prices.

For the **temporal features**, we extracted key variables from transaction and lease commencement dates to represent time trends, depreciation, and short-term market changes. These include transaction year (TX_YEAR), quarter index (QUARTER_INDEX), building age (BUILDING_AGE), remaining lease years (LEASE_REMAINING_YEARS), and derived ratios or decay forms to capture lease effects. To reflect short-term dynamics, we added the 3-month rolling price growth by town (ROLLING_PRICE_GROWTH_3M_TOWN) and the time gap since the last transaction in the same block (MONTHS_SINCE_LAST_TX_BLOCK).

For the **economic features**, we incorporated indicators from public datasets to represent inflation, interest rates, consumer confidence, and market sentiment. These include COE average premium and change rate, CPI value and growth rate, SORA 3-month average rate, STI monthly mean and last closing price, and quarterly GDP growth rate.

All features were standardized to a monthly frequency and merged with the main dataset using the month field. Before model training, they were normalized or standardized to ensure consistent scales across variables.

3) Correlation Analysis and Interpretation: To better understand how temporal and macroeconomic factors affect the housing price per square meter (`PRICE_PER_SQM`), we computed both Pearson and Spearman correlation coefficients for all constructed features and visualized the overall results. This analysis helps identify which variables are most closely related to price variations and provides insights into the underlying market dynamics.

The results for temporal features show clear relationships with housing prices. *TX_YEAR* and its standardized form (*TX_YEAR_STD*) have the strongest positive correlations (around 0.6), indicating a consistent upward trend in resale prices over the years. This pattern reflects the gradual appreciation of Singapore's HDB market. The quarterly index (*QUARTER_INDEX*) also shows a moderate correlation (around 0.4), further confirming the temporal trend in pricing. In contrast, *BUILDING_AGE* exhibits a negative correlation ($r \approx -0.23$), while lease-related features such as *LEASE_REMAINING_YEARS*, *LEASE_RATIO*, and its derived forms (*LEASE_RATIO_LOG*, *LEASE_RATIO_SQRT*) all have similar positive correlations. These findings align with intuition: newer flats or those with longer leases command higher unit prices. Monthly cyclical indicators (*TX_MONTH_SIN* and *TX_MONTH_COS*) show almost no correlation, suggesting weak seasonality in HDB resale activity. Short-term market indicators such as *ROLLING_PRICE_GROWTH_3M_TOWN* and *MONTHS_SINCE_LAST_TX_BLOCK* also have small correlations, implying that localized short-term trading activity has limited impact on long-term price levels.

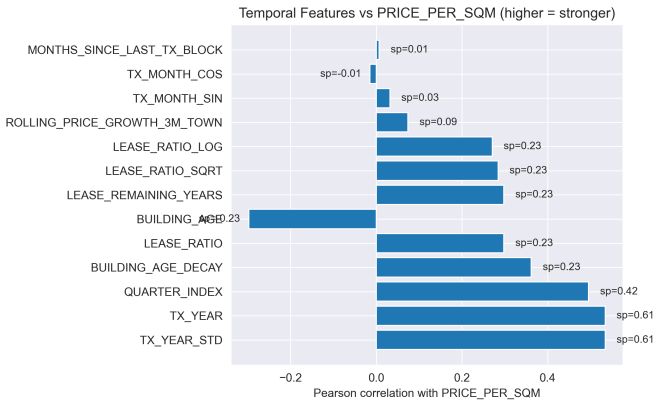

Fig. 9. Correlation between temporal features and `PRICE_PER_SQM`. Blue bars show Pearson correlation, and labels indicate Spearman coefficients.

For economic variables, *CPI_VALUE* and *coe_avg_premium* show the strongest positive relationships with housing prices (both around $r = 0.5$). This indicates that higher consumer prices and higher COE premiums—often associated with stronger economic activity and higher living costs—are accompanied by higher housing prices. The three-month average interest rate (*sora_3m_mean*)

also shows a moderate positive correlation ($r \approx 0.38$), suggesting that housing prices remain resilient during interest rate hikes, possibly due to lagged market reactions. Inflation growth (*CPI_GROWTH_RATE*) and GDP growth (*GDP_Growth_Rate*) exhibit mild but consistent positive correlations, highlighting the link between economic expansion and gradual property value increases. On the other hand, stock market indicators (*STI_Close_Mean* and *STI_Close_Last*) have weaker correlations (around 0.25), indicating limited short-term interaction between equity and real estate markets. Finally, the *coe_change_rate* feature shows almost no correlation, suggesting that short-term COE fluctuations do not directly translate into immediate housing price changes.

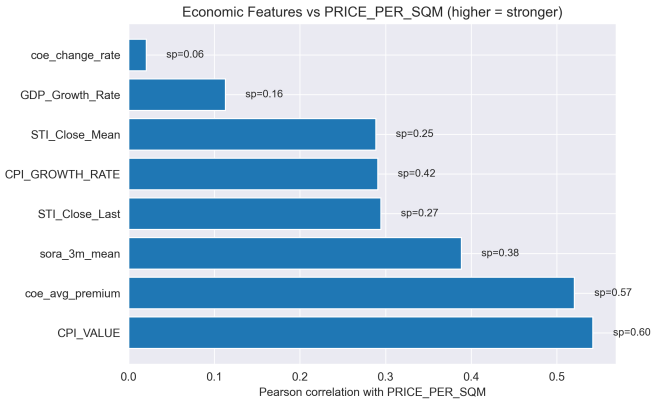

Fig. 10. Correlation between economic features and `PRICE_PER_SQM`. Blue bars show Pearson correlation, and labels indicate Spearman coefficients.

Overall, temporal features capture the long-term market trend and structural depreciation effects of flats, while economic variables reveal how macroeconomic conditions and financial factors indirectly shape housing demand. Features like transaction year and lease duration explain the time-based structure of prices, whereas CPI, COE, and SORA reflect how inflation, consumer sentiment, and borrowing costs influence property values. Combining both dimensions in the model improves predictive stability and provides interpretable insights into how Singapore's housing market responds to changing economic conditions.

III. Data Mining Methods

Our objective is to select the best-performing regression model and optimize its parameters to achieve optimal performance on the validation set, and generalize as much as possible to the test set.

A. Model Selection and Rationale

We selected three tree-based regression models—XGBoost, LightGBM, and CatBoost—all recognized for state-of-the-art performance in regression tasks. Each offers distinct advantages:

- **XGBoost** integrates L1 and L2 regularization directly into its objective function, effectively controlling overfitting. It also supports automatic missing value handling and parallel processing.
- **LightGBM** uses histogram-based algorithms and a leaf-wise growth strategy, enabling faster training while maintaining accuracy, with built-in support for categorical features.
- **CatBoost** specializes in categorical feature processing through ordered boosting, which prevents target leakage and reduces overfitting, even with high-cardinality categorical variables.

These three gradient boosting variants complement each other in regularization, efficiency, and categorical handling, making them suitable for modeling mixed-type features in our dataset.

B. Hyperparameter Tuning and Optimization

We adopted a systematic hyperparameter optimization strategy combining early stopping with Bayesian optimization:

1) An adaptive learning rate and early stopping mechanism were applied during 5-fold cross-validation. Training halted if no improvement in validation loss occurred for 300 consecutive rounds, automatically determining the optimal number of boosting rounds.
2) We used the Optuna framework for a 50-iteration Bayesian search over parameters such as tree depth, subsample ratio, and regularization terms, ensuring efficient and stable model configuration.

This approach balanced exploration of the parameter space with computational efficiency, improving generalization across different data segments.

C. Model Stacking and Ensembling

We evaluated three ensemble strategies:

- **Simple Averaging**: Arithmetic mean of predictions from all three models.
- **Weighted Averaging**: Linear regression on validation set outputs to assign optimal weights.
- **Meta-Learning**: Predictions from base models served as features for a meta-learner (linear regression).

While both weighted averaging and meta-learning achieved lower RMSE on the validation set, they increased test RMSE by 1.32% and 0.72%, respectively, indicating overfitting to validation patterns. Therefore, we selected **simple averaging** as the final ensemble method for its superior generalization performance.

IV. Evaluation and Results Interpretation

A. Performance Comparison

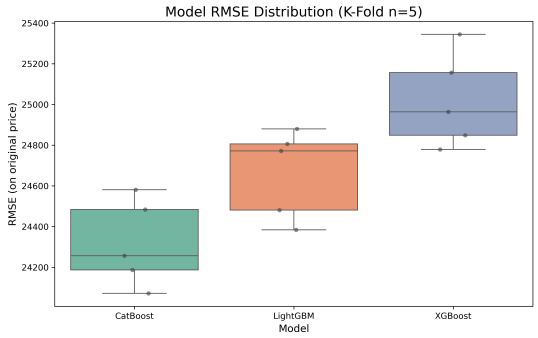

Fig. 11. Model RMSE Distribution (K-Fold n=5)

Our primary performance evaluation, based on the 5-fold cross-validation RMSE, is summarized in Fig. 11. The results show a clear performance hierarchy among the three selected gradient boosting models.

1) **Model Ranking (Accuracy and Stability):** CatBoost emerged as the top-performing model, achieving the lowest median RMSE (approximately 24,260). Furthermore, its boxplot reveals the tightest interquartile range (IQR), indicating the highest stability and consistent performance across the validation folds. LightGBM followed as the second-best model, with a slightly higher median RMSE (approximately 24,780) and a larger IQR. XGBoost was the least performant, displaying both the highest median RMSE (approximately 24,980) and the widest variance, suggesting it was the least accurate and least stable model for this task.
2) **Interpretation (Handling Categorical Features):** This performance difference can be largely attributed to how each model handles categorical features. Both CatBoost and LightGBM provide native support for categorical variables, allowing them to operate directly on our 27-feature dataset. In contrast, XGBoost lacks this native support, which necessitated explicit pre-encoding.
   Despite our efforts to use Target Encoding for high-cardinality features, the pre-processing still resulted in an inflated feature set of 50 features for XGBoost. This dimensionality expansion is the likely reason for its poorer performance, as it increases model complexity and potentially hinders its ability to generalize.
3) **Performance Trade-off (Time vs. Accuracy):** While CatBoost was the most accurate, this came at a severe computational cost. In our experiments, its training time was over 10 times longer than that of the other two models. This positions LightGBM as a highly compelling alternative, offering performance that is competitive with CatBoost while retaining high computational efficiency—a strong balance between accuracy and speed.

B. Feature Importance and Interpretation

To understand the drivers of HDB resale prices, we analyzed the feature importance scores from our models (Fig. 12, 13, 14). While the specific rankings vary, all three models reached a consensus on the key categories of predictive features.

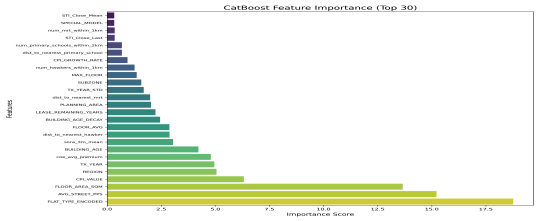

Fig. 12. CatBoost Feature Importance (Top 30)

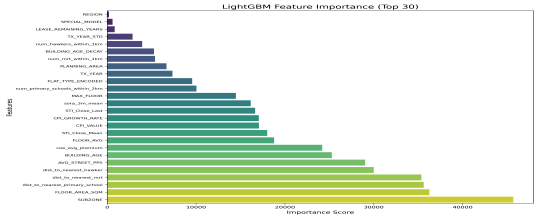

Fig. 13. LightGBM Feature Importance (Top 30)

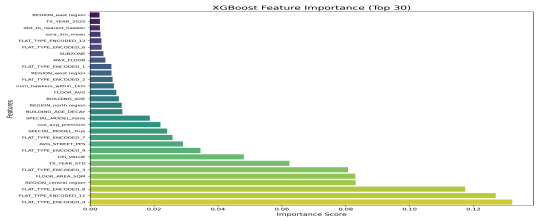

Fig. 14. XGBoost Feature Importance (Top 30)

Across all models, the most significant predictors can be grouped into three categories:

- **Intrinsic Properties:** Core flat attributes like `FLOOR_AREA_SQM`, `FLAT_TYPE_ENCODED`, and `BUILDING_AGE`.
- **Location & Amenities:** Granular location data proved critical. Our engineered feature `AVG_STREET_PPS` (street-level price proxy) and proximity features (`dist_to_nearest_mrt`, `dist_to_nearest_hawker`) were highly ranked.
- **Economic & Temporal Factors:** Market conditions were key, with `TX_YEAR` (transaction year), `CPI_VALUE` (inflation), and `coe_avg_premium` (consumer confidence proxy) all showing significant predictive power.

The differences between the plots highlight the models' distinct architectures:

- **XGBoost (Fig.14):** Shows a fragmented importance list (e.g., `FLAT_TYPE_ENCODED_4`,

FLAT_TYPE_ENCODED_11). This is a direct result of the one-hot encoding it requires, making high-level interpretation difficult.

- **CatBoost (Fig.12):** Offers the clearest interpretation. Its native categorical handling allows it to rank FLAT_TYPE_ENCODED and AVG_STREET_PPS as the clear top-2 features.
- **LightGBM (Fig.13):** Places a striking and dominant emphasis on SUBZONE as its number one feature, suggesting its algorithm found this granular location split to be the most effective initial way to reduce variance.

Overall, the analysis confirms that HDB price is a complex interplay of the property itself, its specific location, and the wider economic environment. The high importance of our engineered, geographic, and economic features validates the significant effort spent on data integration and feature engineering.

C. Error Analysis

An analysis of the out-of-fold (OOF) predictions was conducted to evaluate model performance and identify systemic biases.

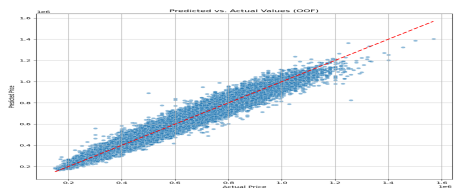

Fig. 15. Predicted vs. Actual Values (OOF).

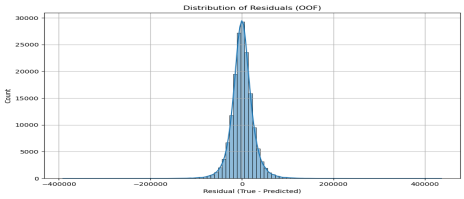

Fig. 16. Distribution of Residuals (OOF).

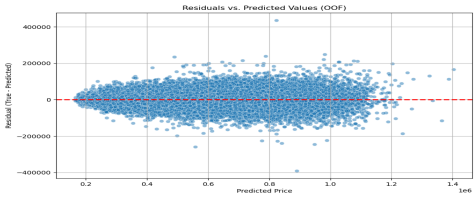

Fig. 17. Residuals vs. Predicted Values (OOF).

The "Predicted vs. Actual Values" plot (see Fig. 15) demonstrates strong overall performance. Predictions are tightly clustered around the $y = x$ diagonal line, indicating a high degree of accuracy for the majority of transactions. Furthermore, the "Distribution of Residuals" (Fig. 16) confirms the model is well-calibrated and unbiased, as the errors form a sharp, symmetric, zero-centered bell curve.

However, the "Residuals vs. Predicted Values" plot (Fig. 17) reveals the model's primary weakness: clear **heteroscedasticity**. The residuals exhibit a distinct "funnel" shape, where the variance of the error (the vertical spread) increases significantly as the predicted price increases. This indicates that while the model is highly precise for low- and mid-priced flats, its predictions for high-value properties (e.g., those exceeding $1 million) are subject to a much larger margin of error.

This issue is likely compounded by the temporal distribution shift noted during the EDA. The dataset contains a higher density of transactions from 2021 onwards, a period which also saw a strong, consistent rise in PRICE_PER_SQM. The model's increased difficulty in predicting these more expensive, recent properties manifests as this amplified variance at the high end of the price spectrum.

D. Limitations and Future Work

This study has several limitations. The model exhibits heteroscedasticity, with higher prediction errors for high-value properties. It is also influenced by temporal data shift, being trained predominantly on recent, rising-market data. Furthermore, granular attributes like unit-specific condition were unavailable.

Future work should focus on enriching the feature set with more detailed property characteristics and exploring advanced models to better capture non-linear relationships, particularly for premium housing segments.

V. Conclusion

This project successfully built and evaluated a machine learning framework to predict HDB resale prices using both intrinsic and external factors. Through careful feature engineering and integration of temporal, spatial, and macroeconomic data, the model achieved stable and interpretable results. Among the tested models, CatBoost demonstrated the highest accuracy, while LightGBM offered strong efficiency and practical value. The findings highlight that location, flat characteristics, and economic conditions jointly shape Singapore's HDB resale market.

Future work could include adding more fine-grained property information (e.g., renovation status, floor plan) and exploring deep learning or hybrid models to capture complex non-linear patterns, especially for premium housing segments.