Unlocking Insights: Predicting HDB Rental through Comprehensive Data Mining

Han Chen
chenhan@u.nus.edu

Ruonan Yu
ruonan@u.nus.edu

Boyang Cao
e1143584@u.nus.edu

Anjie Jin
e1124511@u.nus.edu

Abstract—**Renting is a topic of great concern for many, and rental prices consistently attract widespread attention. Our primary focus is on the rental market in Singapore, particularly due to the significant influx of foreign residents in recent years, leading to a substantial surge in the demand for rental accommodations. Our objective is to construct a predictive model for rental prices through a comprehensive approach that encompasses an comprehensive exploration of historical basic rental data, an evaluation of essential living amenities, an assessment of economic indicators, and an analysis of other pertinent factors. Also, we adopt FLAML, a fast, lightweight, auto-ml library to aid in parameter searching, thereby enhancing the model fitting and overall performance.**

Keywords—Data Mining, Regression, Gradient Boosting, Hyperparameter Search, Parallel Computing

I. Task Description

The project aims to predict HDB flat rental rates in Singapore by analyzing historical data, economics, and infrastructure. It is geared towards aiding renters and property stakeholders in informed decision-making and strategic property marketing.

II. Exploratory Data Analysis

A. Get First Insights into Basic Data

The dataset includes diverse details about HDB rentals in Singapore, covering location, property features, and monthly rent. Our task is to predict the monthly rent using this information, which falls into categorical and numerical types. Also, we analyze spatial and temporal connections and focus on understanding attribute distribution to extract initial insights.

Category Attributes

Distribution Visualization and Analysis

Initially, we explore data distribution and create visual representations (Fig. 2-1 to 2-4) to grasp data characteristics intuitively. Here, 'furnished' values are uniformly labeled 'yes,' charts are omitted in this section.

In our visualization approach, we used pie charts to analyze categorical data, providing a clear representation of data distribution. Analyzing the dataset distribution yields valuable insights. In reviewing Fig. 2-1 to 2-4 visualizations, we found balanced distribution in location-related attributes like town and region. However, attributes related to flat_type and flat_model displayed significant imbalances. This imbalance might lead to larger prediction discrepancies for less common flat_types and flat_models during forecasting. Considering the uniform geographical spread, a strategic move could involve creating distinct models for each region, potentially enhancing prediction accuracy.

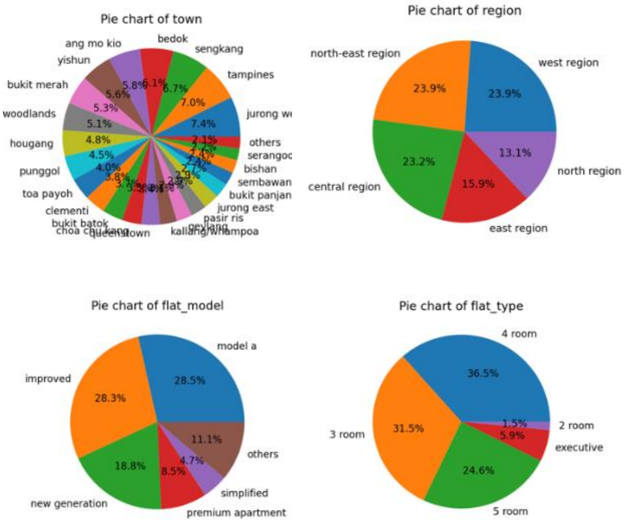

Fig.2-1 to 2-4. Visualization of Categorical Data Distribution

monthly_rent Distribution with Categorical Data

We analyzed the monthly_rent distribution across various categorical attributes and visualized the results using boxplots (Fig. 2-5 to 2-8). Surprisingly, regions seemed to have an insignificant impact on rental prices. However, a significant finding emerged: the '3 room' and '3-room' flat types, though appearing similar, demonstrated distinct distributions. This dissimilarity might arise from differing data sources, potentially representing 'high-end' or 'low-end' HDB categories.

To address this, we plan to fit the data considering both scenarios—one accounting for the '-' differentiation and the other not. The inclusion or exclusion of the '-' may impact data representation. Insights from our Kaggle submissions indicate that resolving the '-' disparity improves predictive modeling outcomes. We acknowledge the possibility of an influence that cannot be immediately dismissed due to our speculation about differing data sources. Therefore, our approach involves modeling the data while considering both scenarios—addressing the '-' issue and not addressing it. We maintained both instances as separate features, allowing the model to autonomously adjust and fit the data according to each case.

Additionally, we have noted a balanced distribution of monthly rent in geographic locations (e.g., region and town), whereas property type categories (model and type) show significant inconsistencies. This is an initial observation, with a more detailed explanation to follow in the 'spatial distribution' section.

Numeric Attributes

Distribution Visualization and Analysis

Alongside categorical feature examination, we've analyzed numerical attribute distributions. Initially, we treated 'approach_date' and 'lease_commence_date' as timestamps. However, 'lease_commence_date' seems more categorical (e.g., houses built in the 80s). Fig. 2-9 highlights non-continuous house construction times, seemingly built in clusters. The distribution matrix (Fig. 2-9) outlines these numerical attributes. Importantly, we excluded the 'elevation' feature as all values are '0'.

As shown in Fig 2-9, 'floor_area_sqm' exhibits 5 peaks, matching the number of 'flat_type' categories, suggesting possible redundancy between the two attributes. Moreover, 'rent_approval_date' demonstrates a relatively even distribution with consistent data volume per month, enabling a similar analysis approach for categorical features. We can model the data separately by year or month. Detailed experimental specifics and result analysis will be extensively covered in the subsequent section.

monthly_rent Distribution with Numerical Data

In Fig. 2-9, 'floor_area_sqm', 'rent_approval_date', and 'monthly_rent' show a clear positive correlation, aligning with our expectations. Larger floor areas, newer lease commencements, and recent rental approvals tend to correlate with higher monthly rents, confirming the assumption that these factors notably influence rental prices.

Additionally, we're focusing on thoroughly exploring the connection between 'rent_approval_date' and 'monthly_rent', detailed in a dedicated section titled 'Temporal Variance'.

Assess Data Quality

Missing Values

All tables, including the primary one (train.csv) and auxiliary tables, show no missing values. This solid data integrity streamlines preprocessing, allowing a smooth transition into exploratory data analysis and model development.

Outliers

Observations from Fig. 2-5 to 2-9 reveal numerous outliers in the dataset, significantly impacting data analysis and modeling. These outliers can bias model parameters, leading to overreactions and affecting adaptability to new data. Initially, we used a basic approach, employing Scipy z-score measure to identify and exclude outliers across all features. However, this lacks depth and scientific rigor.

Realizing the need for a more sophisticated process, we noticed that outlier distribution varies over time intervals, not uniformly across features. In subsequent experiments, we addressed outliers systematically for each feature across time intervals. This strategy aims to align data with rational distribution patterns. Our threshold is set at 3 for z-score values, considering anything greater as an outlier.

Spatial and Temporal Distribution

Upon examining the core data, we noted two crucial feature categories: spatial distribution and temporal data. In practical terms, these data types significantly influence housing prices. During our initial data analysis, we found numerous data attributes strongly linked to spatial and temporal elements, potentially exerting indirect influences. To address both time and location impacts, we have opted to conduct separate analyses for these data categories.

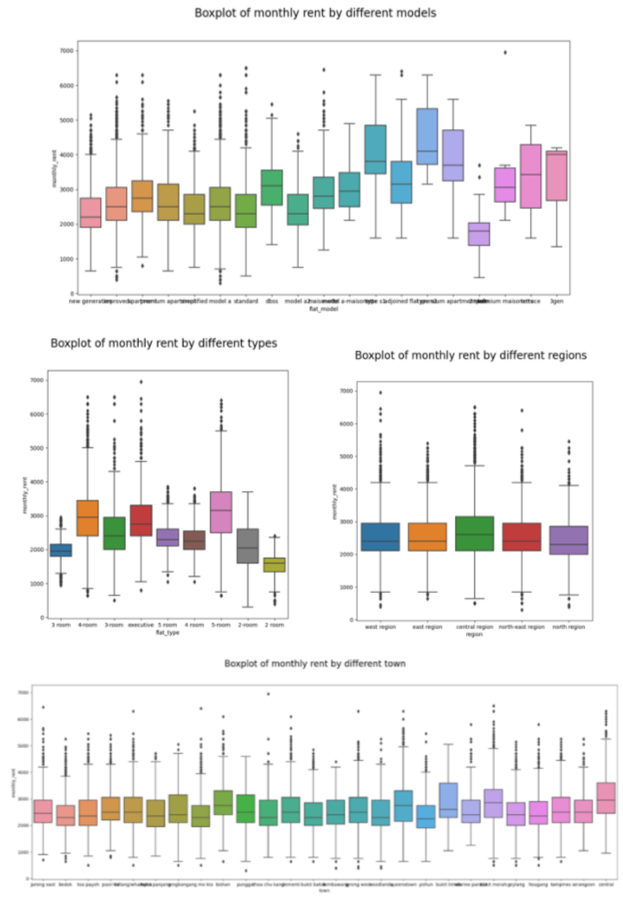

Fig. 2-5 to 2-8. monthly_rent Distribution with Categorical Data

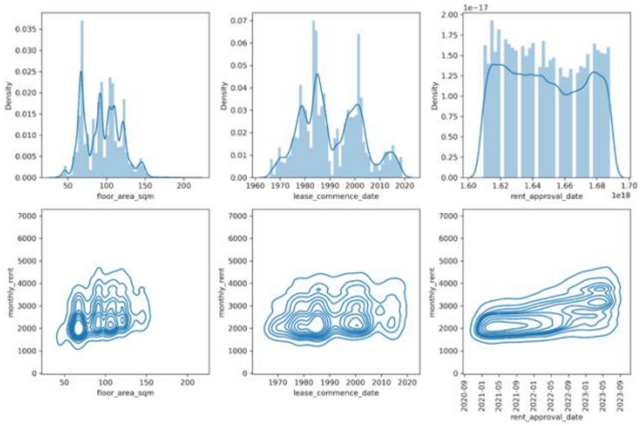

Fig. 2-9. Distribution of Numerical Features and KDEs with rent price

Spatial Distribution

In our prior analysis, we preliminarily explored the relationship between geographic location information (like region, block, and other labels) and the distribution of monthly rent. Observing various geographic labels revealed a relatively consistent overall average distribution of monthly rent.

Moreover, we have introduced more visually intuitive graphs (Fig. 2-10), indicating similar prevalent rental prices in each region, with comparable quantity distributions. However, specific areas display notably high housing prices. Therefore, we're advancing to visualize and analyze the data more granularly, using finer details such as longitude and latitude to deepen the analysis.

We utilized K-Nearest Neighbors (KNN) on the entire training dataset. Initially, we visualized predictions with K=100 in a 3D surface plot (Fig. 2-11), revealing significant price variations across regions, with certain areas showing notably higher prices.

To understand how location impacts rental prices, we trained KNN with the full dataset and used its predictions as a new attribute, replacing other geographic features. Surprisingly, the KNN model's output seems promising for further modeling. This discovery has prompted a focus on geographic attributes, exploring the creation of additional features by combining spatial models like K-means and KNN. This aims to enhance predictive capabilities and understanding of geographic influences on rental prices.

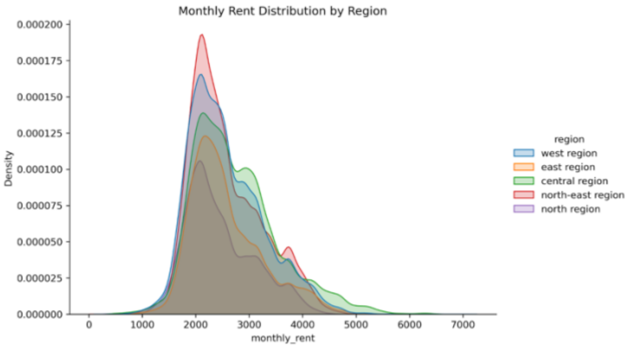

Fig. 2-10. More Visualization of Distribution with monthly_rent and region

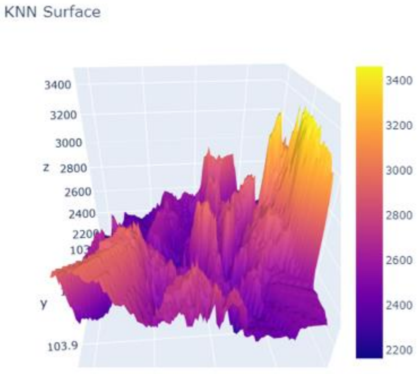

Fig. 2-11. 3D KNN Surface

Temporal Variables

The primary time-related feature in our dataset is the 'rent_approval_date.' To gain a detailed understanding of this, we meticulously plotted the distribution of 'monthly_rent' across various 'rent_approval_year' categories (Fig. 2-12). This analysis revealed substantial differences in rental price distributions across different leasing years.

The findings indicate significant disparities in rental price distributions across distinct years. This suggests that different years may prioritize varying criteria in assessing rental prices, potentially emphasizing factors like floor area or geographic location differently. Additionally, years might be influenced by broader external factors like policies and economic conditions. Notably, we found no evident differences concerning months.

Therefore, we adopted a nuanced approach, creating models capable of interpreting these unique annual distributions. This strategy aims to offer a more comprehensive understanding of evolving factors influencing rental property values over time.

Furthermore, we delved into monthly breakdowns (Fig. 2-13), observing increases and considerable fluctuations in mean, standard deviation, and median rents over time. Although attempting separate models for each month with a small sample distribution of around 2000 did not yield satisfactory results, standardizing monthly rents by month led to desirable effects. This led to a shift in RMSE convergence, improving outcomes. After processing categorical attributes based on standardized monthly rents, reverting to the original values and subtracting the mean value of each month resulted in further improvements without affecting RMSE calculations.

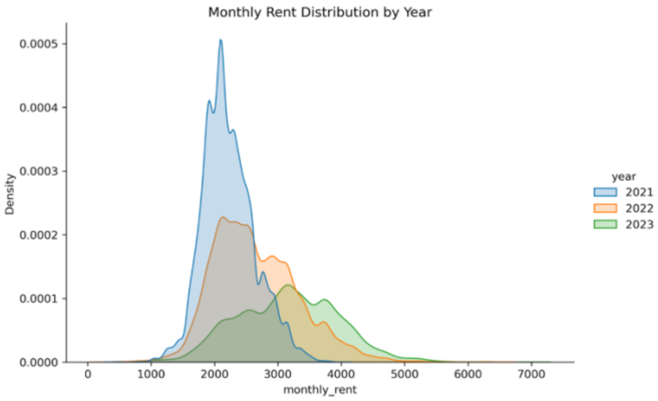

Fig. 2-12. Distribution of monthly_rent via rent_approach_year

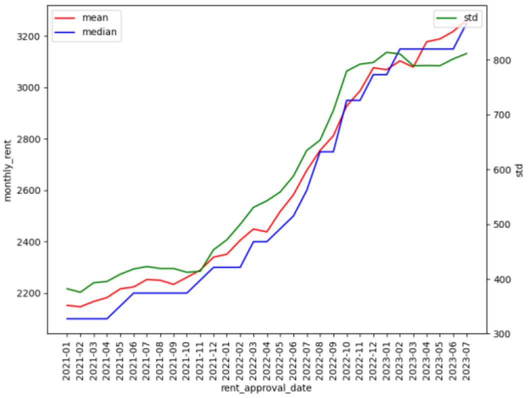

Fig. 2-13. Mean, Median and std of monthly_rent via rent_approval_month

B. Data Preprocessing

Data Cleaning

Fig 2-6 shows 'flat_type' with two categories: '3 room' and '3-room'. Despite their similarity, they display distinct distributions, likely due to data from different sources, representing 'high-end' or 'low-end' HDB properties. To address this, we plan to model the data, considering both scenarios—addressing the '-' issue and not addressing it. We kept both as separate features for the model independent adaptation.

The 'street_name' attribute had variations like 'xxxdrive' and 'xxxDrive'. To ensure consistency and data integrity, all street names in this column were converted to lowercase. Additionally, the 'block' attribute does not uniquely identify a specific housing estate; a block number can be found across various streets nationwide. To precisely associate blocks with their neighborhoods, a combination of 'street_name' and 'block' is necessary. By considering both attributes, we can accurately determine locations and distinguish one housing estate from another.

An additional challenge we encountered is the sparsity within the 'block' attribute, which can potentially lead to overfitting in the subsequent prediction phase. To mitigate this issue, we implemented a threshold system. This involved categorizing less frequently occurring block types as 'others'. Consequently, this approach enables the model to handle block types that might be present in the test dataset but are not adequately represented in the training dataset. Ultimately, this strategy aims to enhance the ability of models to generalize effectively.

Additionally, we have observed the existence of data where all attributes are the same, with only varying rental prices. Our approach to handling this is to consolidate these data points by averaging the 'monthly_rent' value while retaining the consistent attributes.

Transform of Category Attributes

In preprocessing categorical attributes, we initially used target encoding with mean 'monthly_rent' to represent attributes, replacing missing categories with the overall mean 'monthly_rent'. One-hot encoding was avoided due to high category numbers, which could cause dimensionality issues. As no clear ranking order existed for these values, label encoding was unsuitable. Notably, the construction date was treated as a categorical attribute.

However, this approach did not fit the model effectively, showing deviations between predictions and actual distributions in validation data. Consequently, we experimented with various target encoding methods, including standard deviation (std) and median representations, along with normalization and other preprocessing methods. Evaluation using root mean square error (RMSE) revealed that standard deviation (std) improved data distribution fitting.

When using Kernel Density Estimation (KDE) plots to explore data distribution under various target encoding strategies, the standard deviation (std) exhibited the most reasonable distribution representation (Fig. 2-14). Additionally, Fig. 2-15 to 2-16 revealed dispersed blocks with identical identifiers, leading to the modification of the block attribute to a combined street+block attribute.

The issue of limited data samples in attributes like blocks increased the risk of model overfitting. Initially, introducing Gaussian noise to each aggregated block attribute was attempted to prevent overfitting to sparse categories but didn't yield the desired effect.

Ultimately, categories with fewer than a certain value (currently 16) were set as 'others' with an aggregate variance of 0, providing better outcomes. In the test set, categories present in the test data but absent in the training data were also treated as 'others' to enhance the model generalization capability.

C. Linear Correlation with monthly_rent in Basic Data

Category Attributes

We studied the linear correlation between each categorical feature and the target variable 'monthly_rent'. While linear correlation may not entirely reflect the relationships between parameters, we can derive some insights from it.

Categorical data cannot calculate correlations directly, so we used target encoding, replacing all categorical values. Fig. 2-17 displays correlation for all categorical features with target encoding, needing further experimentation for the best method.

From Fig. 2-17, 'block,' 'street_name,' 'flat_type,' and 'subzone' display relatively high correlations with predicted 'monthly_rent' (threshold: 0.3). Notably, 'street_name' and 'subzone' show a very high correlation of 0.8. This indicates complex interaction effects among these variables, suggesting their relationships with 'monthly_rent' aren't solely direct. Changes in these variables may involve intricate mutual interactions, making analysis challenging and introducing multicollinearity issues. In the initial stages, we selected features with a correlation greater than 0.3. Subsequently, we realized that correlation only reflects linear relationships and doesn't precisely depict relationship strength. Hence, we implemented PCA, gradually reducing dataset dimensionality by iteratively eliminating components based on their contribution to overall variance.

Numerical Attributes

In line with the correlation matrix shown in Fig 2-18, 'rent_approval_date' emerges as the most influential factor affecting the 'monthly_rent' price, surpassing even the impact of 'floor_area_sqm.' This finding significantly deviates from our initial subjective perception. Furthermore, 'lease_commence_date' appears to have a minor effect on the price. It's plausible that in the Singapore rental market, the age of HDB estates may not substantially influence the rental prices.

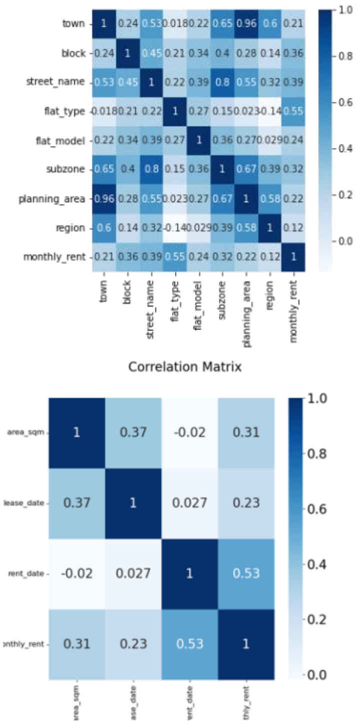

Fig. 2-17 to 2-18. Correlation between Categorical Variables and Numerical Variables and 'monthly_rent'

D. Auxiliary Data

Data on Basic Infrastructure

Overview - Preliminary Analysis

The 'Data on Basic Infrastructure' includes 4 datasets: sg-mrt-existing-stations.csv, sg-mrt-planned-stations.csv, sg-primary-schools.csv, and sg-shopping-malls.csv. These datasets capture info on amenities in Singapore - mall, school, and subway locations. Generally, there's a strong link between rental prices and proximity to these amenities. Typically, residing near convenient amenities correlates with higher rent. Consequently, we conducted an initial analysis to explore this connection.

In this section, we measured distances from the main table to the nearest mall, subway station, and primary school using latitude and longitude data. We excluded data from sg-mrt-planned-stations.csv as planned stations may only have a short-term rent impact, potentially not significantly affecting our analysis. The linear correlation between these distances and "monthly_rent" is displayed in Fig. 2-19. Surprisingly, the graph indicates a minimal effect of nearby facilities on rent prices. This unexpected result might be due to the specific nature of these amenities; for instance, proximity to large malls, busy subway stations, or top-performing primary schools could result in higher rental costs.

Therefore, we conducted experiments to determine which amenities more likely affect rental prices. This involved calculating distances between each property and various amenities to identify those with a significant impact on rental values.

Linear Correlation of monthly_rent and Distance to Primary School

In our initial analysis, we added the attribute 'distance to the closest primary school' to the dataset. However, after computing, we found a weak correlation with monthly rent. This led to a discussion where we hypothesized that the impact on rental prices might differ based on specific schools. We have visualized all locations of the primary schools in the map. We considered that only a few 'famous schools' might significantly influence rental prices in their vicinity. Therefore, we analyzed the relationship between 'monthly_rent' and 'distance to each school' using various transformations like 'd,' '1/d,' and '1/d^2.' Despite our efforts, all coefficients obtained were extremely small, with the maximum value below 0.1, as shown in Fig.2-21. Hence, we shifted our focus to explore alternative features.

Linear Correlation of monthly_rent and Distance to Malls

The mall proximity effect on HDB rental costs is intricate. Our analysis using latitude and longitude shows a positive correlation, confirming that malls enhance convenience and property desirability. Yet, the correlation is weaker than anticipated, with most coefficients below 0.1, as seen in Fig. 2-22. Notably, places like Canberra Plaza, Junction Nine, Northpoint City, Sembawang Shopping Centre, and Sun Plaza reveal slightly stronger impacts, indicating potentially more influence in these areas.

While a positive relationship between rental prices and mall proximity is discernible, its effect remains marginal. Intriguingly, some coefficients even display negativity, reaching as low as -0.05, suggesting lower rents in areas closer to specific malls. This raises pertinent questions about additional influencing factors, such as unique mall characteristics or broader neighborhood attributes.

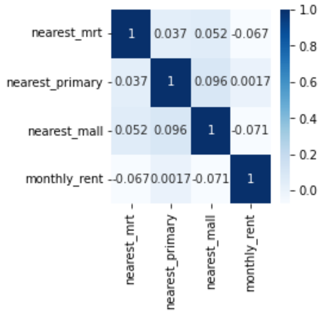

Fig. 2-19. Correlation between Calculated Distances and monthly_rent

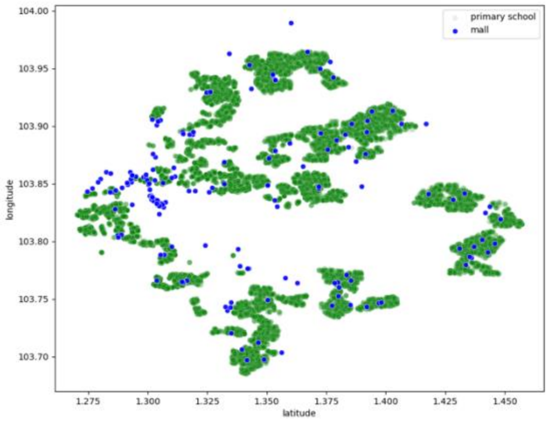

Fig. 2-20. Geographical Location of the Primary School

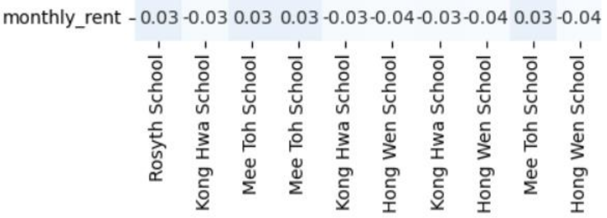

Fig. 2-21. Linear Correlation between Primary School and monthly_rent

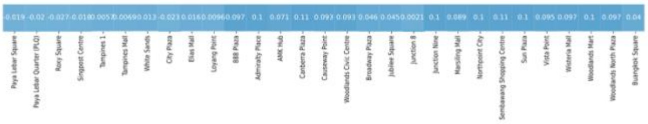

Fig. 2-22. Linear Correlation between Mall and monthly_rent

Linear Correlation of monthly_rent and Distance to Malls

We extended our analysis to evaluate the potential influence of the highly efficient subway system on monthly rental rates. Initial assessments indicated a minimal correlation between rental properties and their proximity to the nearest subway station, defying our initial expectations.

To delve deeper, we identified influential subway stations and conducted an extensive correlation analysis for each station. Even after employing an inverse function to consider distance, the correlation coefficients consistently remained modest, consistently below 0.16, as shown in Fig. 2-23. Our findings suggest that Mass Rapid Transit (MRT) stations may not significantly impact rental rates. This could be attributed to Singapore's efficient public transportation system and the successful implementation of the Certificate of Entitlement (COE) system, effectively addressing traffic congestion concerns.

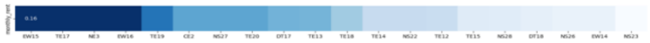

Fig. 2-23. Linear Correlation between MRT and monthly_rent

Data on Objective Economic Environment

The economic environment is indeed one of the crucial factors influencing the rental housing market. For instance, during a downturn in the stock market, landlords might increase rent to maximize their earnings. Conversely, when Certificate of Entitlement (COE) prices are high, signaling a favorable economic situation, rental rates might stabilize. However, these are just speculations, and the specific impacts require further in-depth analysis.

Linear Correlation of monthly_rent and COE

The COE table contains information on year, category, month, bidding price, quota, and bids. Specifically, it documents the monthly transaction prices (bidding price) and the allocated quota for each of the four categories (a, b, c, e), along with the number of bids.

Our initial approach was straightforward. We aimed to calculate the average transaction price for all vehicle types each month, as well as the ratio of bids to quota. We believed that the former could reflect the economic conditions to some extent, with higher transaction prices potentially indicating a better economic environment. The latter could signify the intensity of applications and also provide insights into the economic situation. A higher bids-to-quota ratio indicates a heightened demand for COEs, reflecting a relatively robust economic setting. These conclusions demand a more specialized financial understanding, as our analysis here is relatively basic. The correlation with 'monthly_rent' is depicted in Fig. 2-24. As seen in the figure, a strong positive correlation exists between the average transaction price of COEs and 'monthly_rent,' validating our initial hypothesis. However, the bids-to-quota ratio appears to have minimal predictive effect.

This approach is simplistic considering the diversity of COE vehicle categories; not all represent the entire economic situation. We believe smaller vehicle categories, more tied to daily life, might better reflect the economic environment. To explore this, we calculated separate monthly averages for each vehicle category, covering price, bids, quota, and bids-to-quota ratio. We then examined correlations, as shown in Fig 26. Surprisingly, we found that bids, quota, and average transaction prices for all vehicle categories have a relatively strong positive correlation with 'monthly_rent.' Analyzing the correlation between date and these COE metrics also revealed a significant relationship. This suggests that the strong correlation with 'monthly_rent' might indirectly stem from the strong date correlation, prompting further analysis.

Additionally, we visualized the monthly average price and quota for each type of COE, as depicted in Fig. 2-26. From the figure, it is evident that prices fluctuate noticeably over time, aligning with the trend in housing prices. However, there doesn't appear to be a significant relationship with the quota.

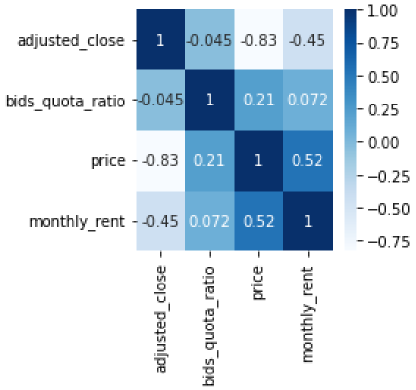

Fig. 2-24. Linear Correlation between COE and monthly_rent

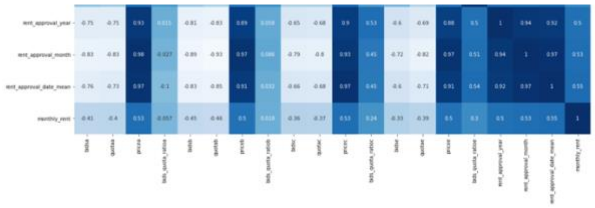

Fig. 2-25. Linear Correlation between COE for Different Types and monthly_rent

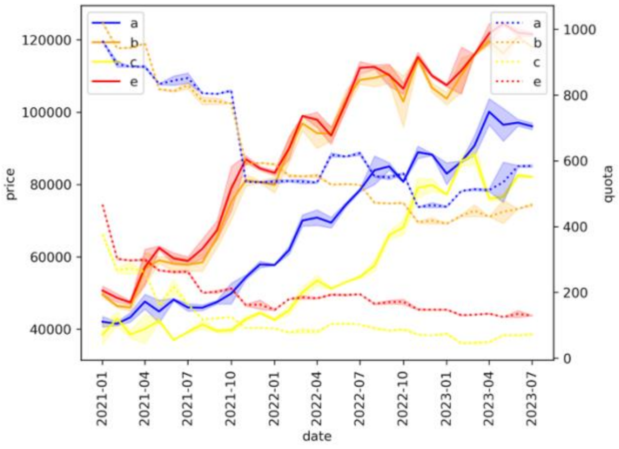

Fig. 2-26. Monthly Average Price and Quota for Each Type of COE

Linear Correlation of monthly_rent and Stock

The 'stock' table records stock information for major companies in Singapore, including attributes such as name, symbol, date, open, high, low, close, and adjusted_close. Due to our lack of expertise, we opted to simply calculate the average adjusted_close for each date as shown in Fig. 2-27. Based on the results, there is a relatively strong negative correlation between adjusted_price and 'monthly_rent,' aligning with our expectations. It is plausible that during a downturn in the stock market, landlords might lean towards raising rents to augment their income. However, it is important to note that even in the results from Fig. 2-27, there remains a very strong correlation between stock prices and 'rent_approval_date.' This raises a speculation that adjusted_close might be indirectly related to 'monthly_rent.' We further visualized the monthly average adjusted close prices for each stock, as shown in Fig. 2-28, revealing a relatively weak relationship with time changes.

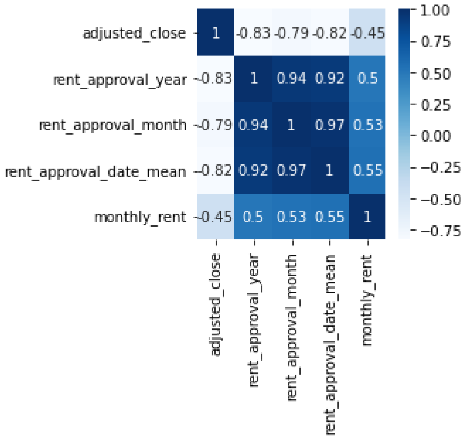

Fig. 2-27. Linear Correlation between Adjusted Close and monthly_rent

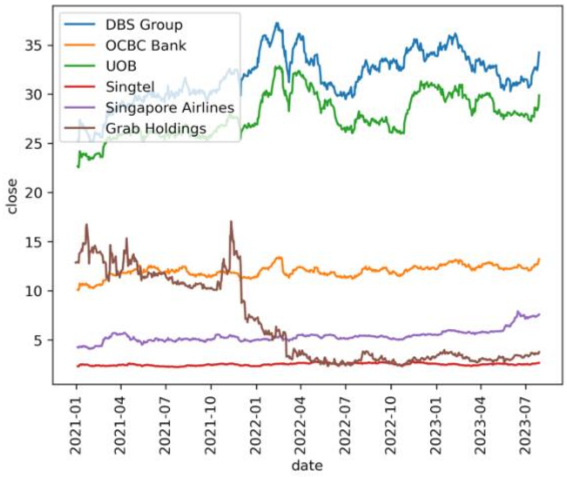

Fig. 2-28. Monthly Average Adjusted Close for Each Type of Stock

E. Additional Attributes / Custom Attributes

Our newly added features primarily focus on geographical information. In the original dataset, there are numerous geographical features such as 'block,' which are often categorical and have a wide range of categories. Directly applying methods like target encoding might result in significant errors, especially for categories with limited data. To address this, we utilized the KNN and DBSCAN methods to generate respective clustering information.

KNN features

To alleviate the impact of geographical locations, our approach involved utilizing standard deviation, which showed better performance in fitting RMSE loss, as mentioned earlier. To generate deviation using KNN, we trained the KNN model with 'monthly_rent' and its square for each K we applied. This process allowed us to compute the variance using the formula $\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$. We generate the 'std-KNN' for K from 16 to 128 to add more detailed geographic information.

DBSCAN Cluster

As mentioned above, the geographic attributes like region are too coarse while the others are too fine. So we generate a category label to separate the locations into clusters smaller than region but bigger than town and area with DBSCAN. Since it can merge the points with density. The DBSCAN result shows in Fig. 2-29.

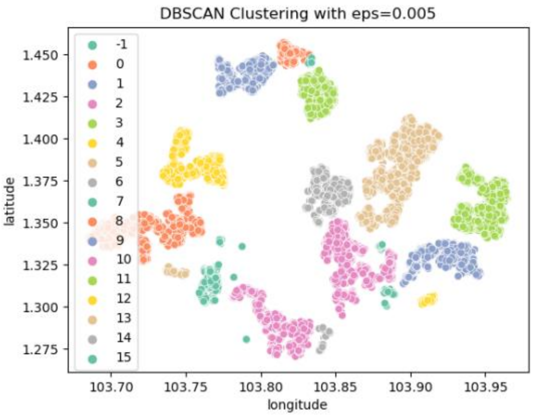

Fig. 2-29. DBSCAN Result

F. Feature Selection

Initially, we used linear correlation for feature selection. However, in later experiments, we found this approach inadequate due to its oversight of non-linear relationships. Subsequently, we employed a top-down PCA method and FLAML, a lightweight library, for additional hyperparameter search and model fitting. Fig. 2-30 shows feature importance.

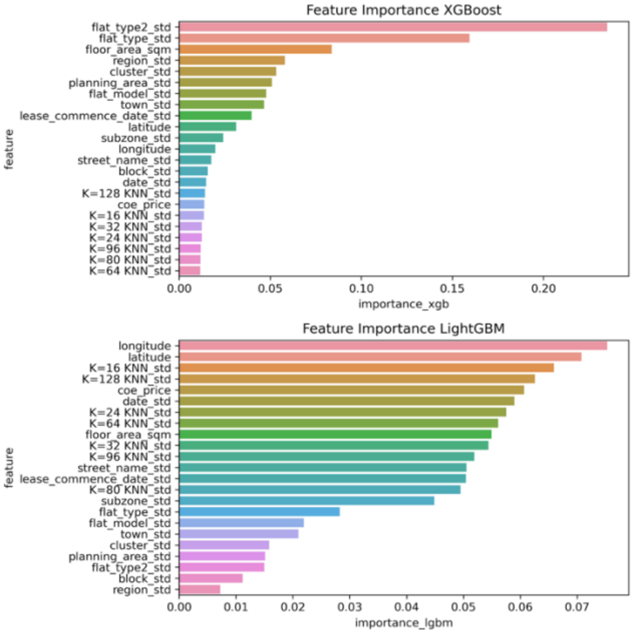

Fig. 2-30. Feature Importance

III. Experiments

A. Main Experimental Procedures

We initially split the training data into a 70:30 ratio for the training and validation sets. This enabled us to assess predictions using validation data and measure performance using metrics like RMSE and MAE. Initially, we explored classic methods such as decision trees, random forests, gradient boosting regression, and fully connected neural networks. Boosting methods showed superior results. As a result, we utilized the top three boosting methods: XGBoost[1], LightGBM[2], CatBoost[3], for further analysis.

Multiple methodologies were implemented to enhance predictive accuracy. We initially explored diverse techniques to process categorical attributes, reduce temporal bias, and investigate auxiliary data. Innovations such as std-KNN and new category features generated through DBSCAN clustering were introduced to increase segmentation granularity.

Following extensive data refinement, our focus shifted to optimizing the models. We utilized the FLAML library from Microsoft for hyperparameters and the Ray library for distributed parallel search on the soc-cluster's xcnf node, leveraging an AMD EPYC 7763 processor and 1TB RAM. While GPU resources weren't essential for weak learners, substantial memory and CPU cores were imperative for parallel searching.

Specifically, separate searches were conducted for LGBM, XGBoost, and CatBoost. For LGBM and XGBoost, we eliminated the original categorical data, retaining only the aggregated standard deviation. However, for CatBoost, we preserved the original categorical attributes along with aggregated features.

Upon evaluating each model's performance based on Kaggle results, we accounted for weights and computed the weighted harmonic mean, geometric mean, arithmetic mean, and square mean. The harmonic and geometric means tended towards lower values, whereas the square mean favored higher values. Significantly, the geometric mean displayed slightly superior performance, indicating a slightly more frequent occurrence of overestimation than underestimation.

B. FLAML: A Fast and Lightweight AutoML Library

FLAML[4] realizes a faster search by taking training time of different hyperparameters into consideration. It introduces the ECI(Estimated Cost for Improvement), which measures the estimated expense for subsequent improvements and starts searching from a low cost hyperparameter point. When the learner is constant, two scenarios arise: (a) discovering configurations with lower loss under the current sample size; (b) increasing the sample size under the current configuration.

As formula below, ECI1 corresponds to situation (a), with K1 and K2 as costs from recent loss enhancements, K0 as the total search cost. ECI2 deals with situation (b) as sample size increases. By controlling the time cost, FLAML can find good parameters faster as well as avoid overfitting with large parameters.

$$ECI_1 = \max(K_0 - K_1, K_1 - K_2), ECI_2 = c \cdot \kappa_l$$

$$ECI = \max\left(\frac{(\tilde{\epsilon}_l - \tilde{\epsilon}^*)(K_0 - K_2)}{\delta}, \min(ECI_1, ECI_2)\right)$$

C. Implementation

We search hyperparameters with settings in table.3-1 for each model with an early stopping. To make a full utility of the 64-core cpu with Hyper-Threading, which set the parallel trials as 32 and each trial we use 4 threads to train the model. So we can utilize all the 128 threads. Since we have a 3-day maximum time for each cluster job, we give each of 3 models a 20-hour time limit and search all the parameters for each model in table.3-2.

TABLE. 3-1. TABLE TYPE STYLES

time limit	threads	parallel trial	K-fold	metrics
20 hours	4	32	10	RMSE

TABLE. 3-2. SEARCHING PARAMETERS FOR EACH MODEL TABLE TYPE STYLES

LGBM	n_estimators, num_leaves, min_child_samples, learning_rate, log_max_bin (logarithm of (max_bin + 1) with base 2), colsample_bytree, reg_alpha, reg_lambda
XGBoost	n_estimators, max_leaves, min_child_weight, learning_rate, subsample, colsample_bylevel, colsample_bytree, reg_alpha, reg_lambda
CatBoost	early_stopping_rounds, learning_rate, n_estimators

With the searching log in figure.3-1 and figure.3-2, we can say that we have found good enough hyperparameters for each model. By ensembling the prediction, we achieved a score of 477.93 on kaggle and rank in top-2, which is a satisfactory result.

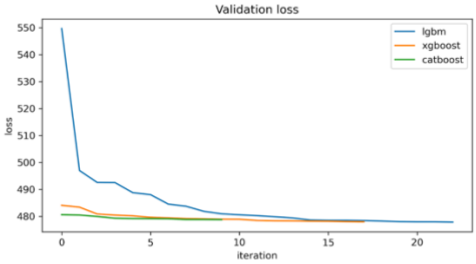

Figure. 3-1. Cross-Validation Loss for Searching Iteration

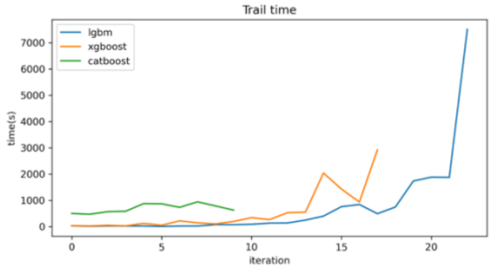

Figure. 3-2. Searching Time for Searching Iteration

REFERENCES

[1] [1] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[2] [2] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[3] [3] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.

[4] [4] Wang C, Wu Q, Weimer M, et al. Flaml: A fast and lightweight automl library[J]. Proceedings of Machine Learning and Systems, 2021, 3: 434-447.

[5] Christian von der Weth. (2023). CS5228-2310 Final Project. Kaggle.https://kaggle.com/competitions/cs5228-2310-final-project