Singapore HDB Rental Price Analysis & Prediction using ML Algorithms

ST4248
Term Paper

Fan Zixian (A0267684L)

Abstract

Housing rental is one of the very significant expenses in people's life. However, rental housing prices will increase significantly in Singapore in 2022, which may seriously affect people's living costs. Therefore, it is imperative to predict the practical significance of rental housing prices accurately. This paper implements different machine learning and deep algorithms to incorporate macroeconomic indicators and housing-specific information to predict the HDB rental price in Singapore. This study also hopes to explore the impact and reasons of different features on the rental price through the obtained better model.

1 Motivation

Housing is one of the essential requirements for people to meet living requirements, and it occupies a significant part of the cost of living. According to a study by the PLB, rents in Singapore soared by about 23-28% island-wide in 2022, which imposed enormous pressure on tenants[1]. In the face of the rising prices of rental properties in Singapore, using quantitative approaches to predict the cost of rentals has become increasingly valuable. Rentals in Singapore can be classified into Housing & Development Board rentals and private rentals. However, since the quality of private rental housing is difficult to standardize, the provision of insufficient data leads to large fluctuations rental price. In addition, the data from URA is also not provided regularly. Hence, this study focused on the modeling and forecasting of HDB prices. According to Bloomberg's research, rental housing prices may correlate with macroeconomic conditions[2]. This study considered the impact of Singapore's economic conditions on rental housing prices and hoped to examine the reasons for the boom in rental housing prices in Singapore in 2022.

2 Data Collection and EDA

2.1 Data Collection and Cleaning

HDB rental prices were sourced from the HDB Rental Statistics website[3]. The statistics provide the median apartment rent for 26 Districts in Singapore from 2007 Q3 to 2022 Q4 as a guide. The apartments include six types, 1 to 5 bedrooms, and finally, Executive. From the data, the types of houses with less than 20 rentals per quarter will not be published on the website, and most rentals with Room types of 1 or 2 people belong to the situation that should not be published. Including them for training may affect the accuracy, so they were excluded from this study. After removing other null values, the data set was constructed and screened with the monthly house rent as the dependent variable. Data observation found a substantial monthly rental price, which was $16,800 higher than the second-largest monthly rent, so the study chose to exclude it, leaving a total of 4777 data.

According to Bloomberg's analysis, the reason for the sharp rise in rental housing prices in Singapore in 2022 might be related to Singapore's continued economic recovery, flexible employment, and household income, and it was expected that there would be more remarkable growth in 2023[2]. These showed that the macroeconomic situation could indeed affect residential rental prices, which makes an impact because the macroeconomy's support comes from the individual's economic situation. Therefore, this study selected several macroeconomic indicators for investigation, all from the Department of Statistics Singapore, from Q3 in 2007 to Q4 in 2022[4]. Note that, the quarterly interest rate was obtained by taking the average of every 3 months each quarter.

The two data sets were matched according to time and merged together, and a 4777×13 data set was obtained. Classified parameters were in Appendix 1:

2.2 EDA

After obtaining the data, the `Time` column was removed currently, and the remaining variables could be distinguished according to quantitative and qualitative. For qualitative variables, the study first adopted the dummy variable processing method to expand all HDB House Specific Information-related variables and finally obtained 30 variables. The four dummy variables of `Room Type` were presented in Appendix 2. Although there was a slight difference in proportion, the number of each type was relatively sufficient, and there was no need for Class Imbalance processing. Observing the Correlation Matrix of the 30 variables, it was detected that the correlation is small. Although there were many variables, it was not clear whether the impact of the `District` on the `Rental Price` was based on its location, economic status, population density, etc. Therefore, it couldn't rely on domain knowledge for feature merging. All qualitative variables would be preserved.

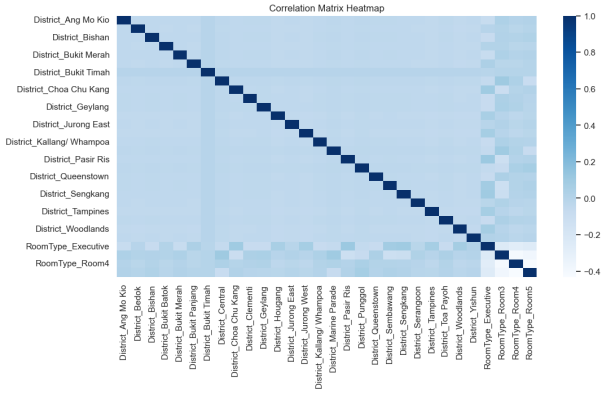

Figure 1: `Correlation for Dummy Variablesl`

For quantitative Macro Economy Indicators', their calculations are generated based on Singapore's economic conditions and have inherent correlation, so the study considered VIF values to measure the collinearity between variables. According to the figure below, many variables had very strong collinearity, hence can consider removing some variables before fitting the model.

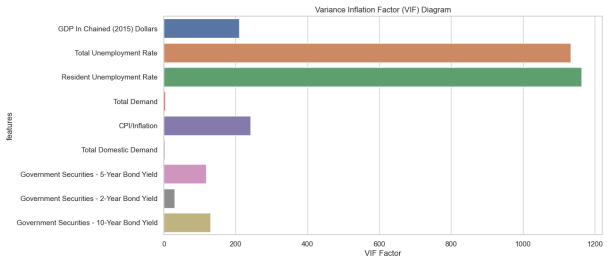

Figure 2: `VIF for Macro Economy Indicatorsl`

For the response variable `Rental Price`, the time attribute was ignored to present its histogram in Appendix 3. It was observed that the data has prominent bell-shaped characteristics and no outliers, so no need to do redundant processing such as log transformation.

3 Data Processing

3.1 Feature Selection

Based on EDA, this study performed feature selection by step-by-step path for Macro Economy Indicators, and finally got features shown below. All of their VIF values were below 15, indicating relatively independent relationships.

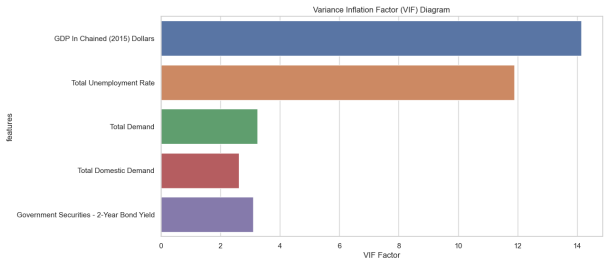

Figure 3: `VIF for Selected Macro Indicatorsl`

3.2 Data Scaling

Since the scales of quantitative variables were quite different (e.g., `GDP` vs. percentile numbers), it was necessary to scale the data for the accuracy of model training and the efficiency of gradient descent. This study considered Min Max Scalar w.r.t. Standard Scalar because these Macro Indicators didn't show strong normal patterns.

4 Model Training, Prediction, and Evaluation

Then, this study merged the quantitative data after scaling with all qualitative data and obtained 4777×35 data sets under two scales. In this study, 80% and 20% ratios were used to split the training and testing data sets. The study would consider the following models according to ideas from others' work[5], with Grid Search Cross Validation tuning:

- Ridge Regression
- Random Forest
- XGBoost
- LSTM

For general models, the research ignored the time attribute. The data matching relationship was to match the current rental price with macroeconomic indicators without lagging, which meant that the actual application of the model needed to estimate the future macroeconomic indicators first (because logically the rental price should be a reflection of the current macroeconomic conditions); for the LSTM model, the research would consider time attributes, because the past time series may also have an impact on the forecast. However, noted that it is possible to give a bad result as most of the data has similar timestamps, which is not suitable for LSTM training. MSE and R^2 were chosen to be the evaluation criteria for different models.

4.1 Ridge Regression

Ridge regression is a regression model where the loss function is the linear least square function. It imposes l2-norm regularization during the fitting process to make the weights of different features more balanced. This study considered performing tuning regarding parameters: alpha, fit_intercept, and solver. Alpha is the most critical parameter of the model, as it is used to set the strength of the penalty, and the best alpha after tuning was 1. The solver is chosen to determine how Ridge Regression calculates the cost function, which affects the model performance to some extent. After tuning, the best solver in this study was "svd", which can calculate the Ridge coefficient by decomposing the singular values of the independent variables and finally provides more stable results[6].

Ultimately, the best model came from the standardized data, which provided a 48002 Test MSE and an R^2

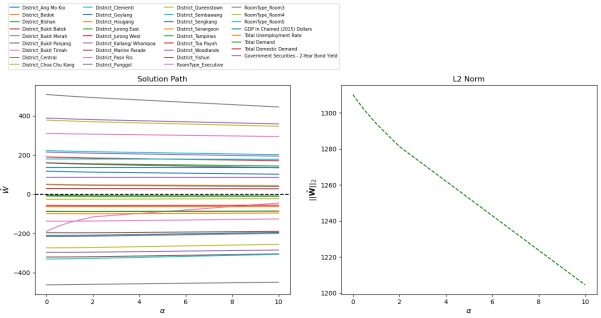

Figure 4: Ridge Regression Regularization Performance

of 0.7212(Table 1). This result didn't perform well, and as seen from the graph above, the increase in alpha does not have a good effect on the model fit, either because the vast majority of the parameters in the model have good significance, or because linearity is not sufficient to describe such a complex relationship.

Table 1: Ridge Regression Model Performance

	Standard Scaler	Min Max Scaler
Train MSE	47775.5627	47778.1146
Train R^2	0.7210	0.7210
Test MSE	48002.9465	48015.2520
Test R^2	0.7212	0.7212

4.2 Random Forest

The random forest model is capable of constructing decision trees by randomly selecting data and feature subsets to reduce the risk of overfitting, and to use the average of the decision tree predictions to obtain the final prediction results. In the random forest fitting process, the parameter set considered both the proportion of feature subsets to be used in each fitting, and the limitation of the depth and complexity of the decision tree. The model performance after parameter tuning is shown in the following table, the best MSE of the test set is 155119, and the best R^2 is 0.9122(Table 2), which is an enormous improvement over Ridge Regression.

Table 2: Random Forest Performance

	Standard Scaler	Min Max Scaler
Train MSE	8172.0496	8359.9200
Train R^2	0.9523	0.9512
Test MSE	15119.1581	15166.8985
Test R^2	0.9122	0.9119

From pruning, the best feature proportion parameter was to consider including full features at each decision tree fitting, making the Random Forest a Bagging model, which indicated that all the features had a high value for the judgment of the results. According to the demonstration of the feature importance graph, `GDP`, `Room Type 3` show the highest value in decision-making, and they can better help in node impurity reduction and model prediction. Other macroeconomic and regional indicators also considerably impact, proving the appropriateness of feature selection.

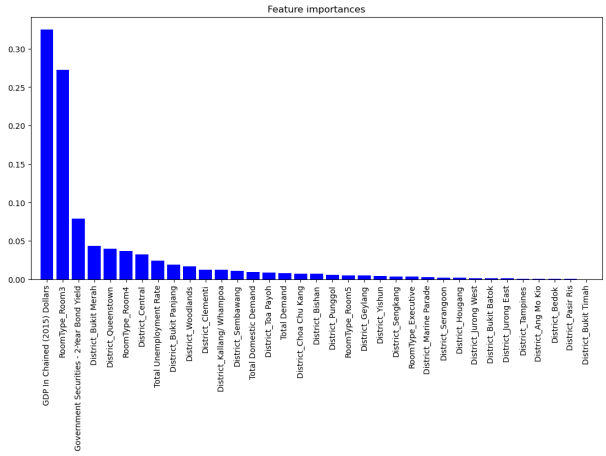

Figure 5: Random Forest Feature Importance Graph

4.3 XGBoost

XGBoost is a compelling Ensemble learning model that improves prediction accuracy by integrating multiple weak decision trees and letting subsequent trees iterate over the gradient based on the residuals of the previous decision trees. In addition, XGBoost can use regularization parameters to prevent overfitting problems, making the prediction results more robust and accurate than other models. The most critical parameters of XGBoost are related to his learning rate and the depth of the set decision tree because it advocates slow learning to improve the fitting effect. At the same time, regularization parameters similar to the elastic net will also be taken into consideration by Grid search. In this study, XGBoost gave the highest result, the MSE of the test data set could be reduced to 5143, and the R^2 could reach an astonishing 0.9701(Table 3), which is also a tremendous improvement compared to random forest.

Table 3: Random Forest Performance

	Standard Scaler	Min Max Scaler
Train MSE	3486.5387	3486.5387
Train R^2	0.9796	0.9796
Test MSE	5145.5533	5143.3276
Test R^2	0.9701	0.9701

Based on the excellent performance of XGBoost, this study focused more on the analysis and interpretation of the XGBoost model. The learning rate in the best parameter was 0.3, which was relatively fast, showing that XGBoost could handle the research data better than other models. After Re-examining the Feature Importance Graph obtained by the XGBoost model and comparing it with Random Forest: the feature with the highest importance value is `GDP`, but the performance of other features in XGBoost is more balanced, and more macroeconomic indicators appear more significant (such as `Total Domestic Demand` and `Total Demand`). These findings show that in XGBoost, more macroeconomic indicators are used to split nodes, which can prevent some features from being misidentified due to high variance in Random Forest[7].

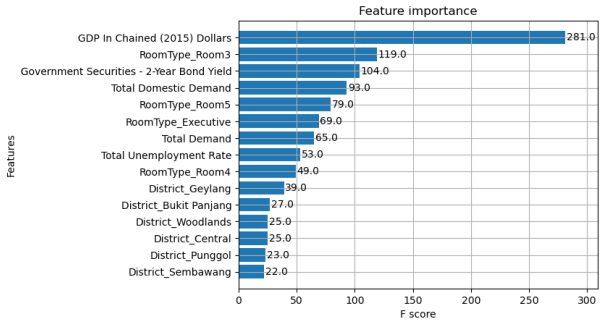

Figure 6: XGBoost Feature Importance Graph

To better explain the XGBoost model, this study adopted a game theory-based comprehensive decomposition and calculated the prediction contribution of different parameters in the test set to test instances and presented in the SHAP diagram. By analyzing the structure of the SHAP graph, the most valuable prediction feature is the GDP variable, which also shows a very obvious variance, which can explain the significant impact of the macroeconomy on rental prices because `GDP` is a very intuitive presentation of Singapore's macroeconomy. All `Room Types` are also shown in the figure, and the larger the Room Type, the higher the rental price, which is also in line with conventional considerations. In addition, `District position` is also a significant factor: similar to more prosperous areas such as Central and Queenstown, the predicted value has a deeper red color and an immense SHAP value, which proves that cities in these areas will be more popular, so the rental price is higher.Note that, if the red point of a feature is only to the left of the center line, it means that the higher value of the feature is associated with the lower predicted value of the model output. This may be due to the linear relationship between the feature and the expected result, or the feature has opposite interactions with other features.

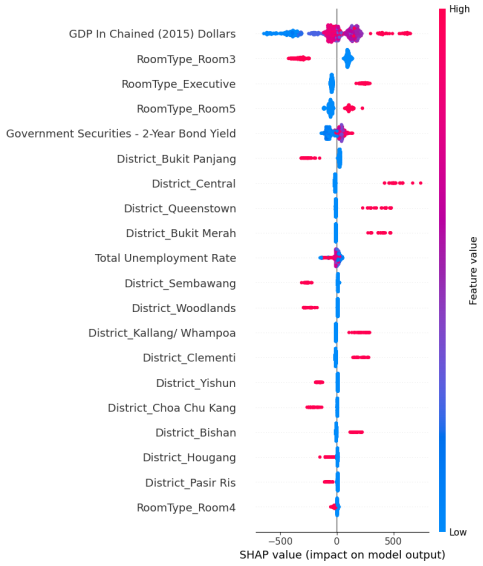

Figure 7: XGBoost SHAP Diagram

However, observing the Tree plot corresponding to a certain decision Tree of XGBoost Appendix 4, it can be noted that its first split is based on `Room Type 3` instead of `GDP`. Reviewing at the SHAP diagram again, it can be seen that the macroeconomic indicators fluctuate around the baseline, indicating that they may not be able to match the rental price completely. This may also be related to the lagging of data because the performance of rental prices may slightly lead or lag behind changes in macroeconomic indicators, and it may be more valuable to consider moving data forward or backward. Therefore, if some district position features can greatly separate the data, they could be regarded as nodes prior to macroeconomic indicators.

4.4 LSTM

LSTM is a variant of the recurrent neural network, which stores long-time information with the help of memory cells and uses input and forgetting gates to selectively adopt past time series information for model fitting, which can avoid the problem of gradient vanishing in traditional RNN models. To make the deep learning model fitting more accurate, this study considered Standard Scaling for `Rental Price` (due to the previously observed bell-shaped distribution of Rental Price), and then inversed back to the original scale for evaluation afterward.

A10-step sequence length was used to create the time series, the batch size was set to 64, and the hidden size and epochs were tuned on GPU without cross-validation for simplicity. However, the results were poor after several training sessions, and the best result was a Test MSE of 174378.89 at 10 epochs, with a R^2 of -0.0123, meaning that the model was even less effective than if it had not been fitted. The problem may be that there is too much data simultaneously in the data and fewer different time series. This would result in a 10-step sequence length containing all data with the same timestamp but various other features; hence the LSTM model could not be fitted for time. And the small amount of data makes it impossible to use a larger step size to incorporate more data at the same time. Improving the effect may require changing a larger dataset and considering how to combine different data so that there are fewer data at the same time node, which may require findings from other models to support the reason for the feature combination.

5 Conclusion and Future Work

According to all the above model training and relevant results, this study finally selected XGBoost as the prediction model, which can provide the best accuracy, Using the different graphs supplied by the model, the study can confirm that several indicators have a reliable impact on rental price prediction.

1. Macro economy indicators: GDP, unemployment rate, government 2-year bond yield, total demand, and total domestic demand, among which GDP is the most influential. Among them, GDP and unemployment rate reflect Singapore's economic conditions. Under the shape of a high GDP growth rate and low unemployment rate, Singapore's economic condition is vigorous, which can support the inflation of rental prices; the yield of bonds can affect people's interest in leasing from the perspective of investment and return; the demand factors may be considered from the overall supply and demand of the Singapore economy, reflecting the impact of supply and demand on housing prices. Therefore, by observing Singapore's macroeconomic indicators in 2022, it is not difficult to find that Singapore is in the recovery stage of the economy, with GDP growth and lower unemployment, so rental prices may rise.
2. HDB House Specific Information: XGBoost finally takes into account all districts and Room Types. The impact of Room Type on rental price is intuitive: the larger the house, the more expensive the rent. In order to study the effect of the district, this study used the obtained best XGBoost model to predict the rental price for Room Type 3 (with the greatest frequency), and all macroeconomic indicators are set to zero.

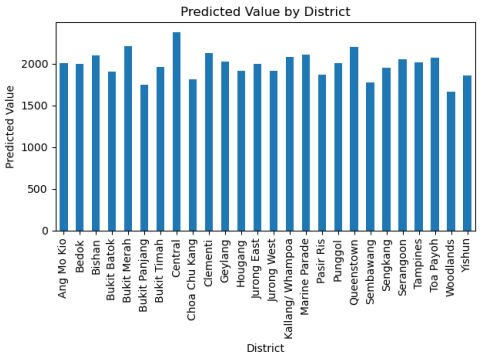

Figure 8: Predicted Rental Price with District differed

Through the analysis of the prediction results, it can be known that in more prosperous residential areas and areas with more convenient transportation, the predicted value of HDB's rental price will be higher. For more remote areas, such as Woodlands, the predicted value of rental price will be lower. More detailed research requires heat map presentation and analysis through GIS charts.

For future work regarding the dataset, the current predicted rent is HDB Rental House, which is the rental supported by the HDB, and it is more controlled and standardized by the government. The data provided by URA or other website APIs can be implemented to consider the private rental housing in the market, and the use value of the research is more extensive. In addition, because the HDB data set is insufficient, it is difficult to incorporate time series attributes into the model's training, so it is impossible to build a good time series deep learning model. Future research can be based on a more comprehensive data set, such as lagging macroeconomic indicators, artificially creating more complex connections. From the perspective of features, there may be more macro economy indicators worth mining that can be considered in the model, and techniques such as genetic algorithms can also be applied to construct new factors.

6 Reference

1. sim, nikki. (2023, January 27). 2022 Singapore Real Estate Market Wrapped: Key Statistics and Trends. PropertyLimBrothers. https://www.propertylimbrothers.com/2022-singapore-real-estate-market-wrapped-key-statistics-and-trends/

2. Bloomberg L.P. (2023, January 16). Soaring Singapore Rents Set to Climb Another 10-15% This Year. Bloomberg.com. https://www.bloomberg.com/news/articles/2023-01-16/soaring-singapore-rents-set-to-climb-another-10-15-this-year

3. HDB. HDB | Rental Statistics. Www.hdb.gov.sg. Retrieved April 11, 2022, from https://www.hdb.gov.sg/residential/renting-a-flat/renting-from-the-open-market/rental-statistics

4. Singapore Department of Statistics. (2022). SingStat Table Builder. Retrieved April 11, 2022, from https://tablebuilder.singstat.gov.sg

5. Ahuja, A., Lahiri, A., & Das, A. (2021). Predicting Airbnb Rental Prices Using Multiple Feature Modalities. ArXiv:2112.06430 [Cs]. https://arxiv.org/abs/2112.06430

6. Scikit-learn: Machine Learning in Python. (2021). sklearn.linear_model.Ridge — scikit-learn 0.23.2 documentation. Scikit-Learn.org.

7. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

7 Appendix

Appendix is for diagrams, and tables that are meaningful to some understandings and inference, but not significant to the logical flow of the main analysis.

Appendix 1:

Time	Rental Price	HDB House Specific Information	Macro Economy Indicators
Won't be considered for training, but can be considered as time series		District: totally 26 RoomType: from 3 to 5 people room type, and	GDP In Chained (2015) Dollars Total Unemployment Rate Resident Unemployment Rate CPI/Inflation Total Demand Total Domestic Demand Government Securities - 5-Year Bond Yield Government Securities - 2-Year Bond Yield Government Securities - 10-Year Bond Yield

Figure 9: `Classified parameters in total`

Appendix 2:

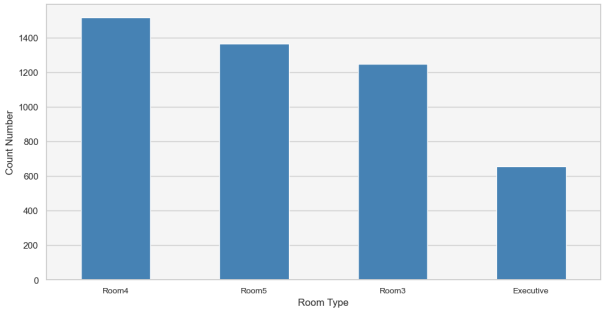

Figure 10: `Room Type Histogram`

Appendix 3:

Figure 11: `Rental Price Histograml`

Appendix 4:

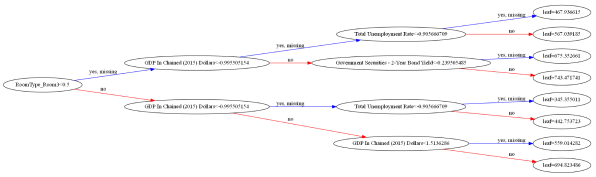

Figure 12: XGBoost Tree Plot