CS3244 Machine Learning: Developing a Model to Predict Future Resale Flat Prices in the Singapore Housing Market

Team 24: Chong Cheng Yang, Lionel (A0202037X), He Cheng Hui (A0190274E), Kwa Jun Liang, Desmond (A0210629L), Ong Zhi Yi, Rachel (A0222167L), Wang Chen (A0221850N)

National University of Singapore (NUS)

e0415846@u.nus.edu, e0325858@u.nus.edu, e0470730@u.nus.edu, e0559699@u.nus.edu, e0559382@u.nus.edu

Introduction

HDB resale flat prices soared in 2021 due to various reasons, including the delay in construction for BTO flats and the constant need for a workspace due to Work From Home arrangements. Such reasons have caused the prices of HDB resale flats to peak. In 2021, the average price per square foot (psf), was $511, higher than the last property peak in 2013[1] of $478 psf. Research conducted by various sources[2] has also shown that even with the latest cooling measures implemented by the government in late 2021, there might not be significant changes in the resale market.

With such uncertainties in the housing market, we hope to come up with a machine learning model that is trained to perform price predictions for homebuyers in Singapore to easily determine how much they should set aside for their dream flat.

Our model leans away from human bias and instead focus more on statistical integrity when predicting prices. Such a prediction will be more objectively representative and reliable compared to market prices listed on property websites which might be affected by the seller's personal factors.

In 2021, first-time home buyers made up about 80 percent of transactions in the current market. It is unavoidable for them to feel overwhelmed by the vast amount of housing information on the internet. This motivated our group to focus on buyers' need to get predictions for resale flats. Our model can help them predict a more accurate "future market price" of a specific resale flat configuration they are looking for.

Applications

Our model aims to mainly help buyers looking for a suitable place in the resale flat sector.

Users of our model can input a series of features they are looking out for in their future house. Ideally, they will be able to input a location, flat type, level of flat, remaining lease and duration to travel to different amenities found in a common neighborhood.

For the current stage of this model, we have decided to focus on training data based on resale flats in Sengkang. In recent years, flats in Sengkang have been gaining popularity[3]. 447 flats were transacted in January and February 2021. According to srx.com.sg[4], Sengkang is one of the towns with the highest traded volume of HDBs over the past 90 days, making it one of the most popular estates.

Sengkang also offers many outdoor recreational options for residents to exercise, multiple public transportation options to connect to other parts of Singapore and retail malls. We hope to focus on the specific amenities within that area, to obtain a more comprehensive prediction. The same research methodology can be applied to different towns in Singapore, should there be a need for future studies.

[1] J, R. (2022, January 8). 6 Major Property Trends To Watch In The Singapore Property Market In 2022. Property Blog Singapore - Stacked Homes. https://stackedhomes.com/editorial/6-major-property-trends-to-watch-in-the-singapore-property-market-in-2022/#gs.x8am7t

[2] Property Giant. (2022, January 4). Cooling Measures in Relation to HDB Resale Prices. Blogs and Market Trends | Property Giant. https://www.propertygiant.com/blog/cooling-measures-in-relation-to-hdb-resale-prices

[3] Kor, V. (2021, March 26). Punggol and Sengkang home to the most HDB resale transactions in 2021 to-date. EdgeProp. https://www.edgeprop.sg/property-news/punggol-and-sengkang-home-most-hdb-resale-transactions-2021-date

[4] Singapore Property Heat Map | Singapore Real Estate and Property Exchange. (n.d.). Singapore Real Estate Exchange. https://www.srx.com.sg/heat-map

Research Methodology

Resale Flat market in Singapore

To understand the current market for resale flats, we interviewed a real estate agent for more information. The interviewee is a Senior Marketing Manager, who has a decade-long experience and has won accolades for her work. Below are some of the factors that buyers tend to consider, which she ranked in order of importance:

1. Credit/Financial ability and Cash Over Valuation (COV)
2. Location
3. Age of the flats
4. Type of HDB (3-Room, Executive, etc.)
5. Floor level
6. Layout of the overall floor plan
7. Direction that the flat is facing
8. Condition of the flat

An article from PropertyGuru[5], further lists some of the important factors that might affect the value of the resale flat:

1. Loans based on the buyer's credit rating
2. Size of the flat
3. Neighbors living nearby - this determines the living condition of the place, and is very subjective to the buyer
4. Proximity of amenities such as MRT stations
5. Condition of the flat

Some other minor factors include where the sun is facing in the flat and how much crime exists in the neighborhood.

Developments for Sengkang started in 1994, when it was then further divided into seven housing estates, developed with an intention to hold up to 95,000 houses[6]. In 2017, over 65,981 HDB units were completed under the Built-to-Order (BTO) Scheme, making Sengkang a relatively new estate.

Additionally, the government has been encouraging Singaporeans to start their own families, to the extent of 'paying' them to have children[7]. To serve a young demographic, large amounts of pre-schools and kindergartens can be found around Sengkang. By 2020, residents there can expect 300 more childcare centres[8]. Resale flats in Sengkang are mostly 4 to 5 room flats, which is to account for the possibility of residents there setting up bigger families.

Existing applications to predict Resale Flat prices

Currently, we found three existing applications of HDB resale flat price prediction, and they are listed below:

1. The HDB One Map[9] shows a detailed visualization of HDB flats in Singapore. However, they lack the ability to filter the HDB flats by features, such as towns or floor area.
2. SRX[4] also provides a visualization of the prices and transaction volumes by towns. However, the prices shown is an average price of the whole town and does not go into details of specific flat prices.
3. Teoalida[10] shows the HDB median resale prices based on towns and flat types. Although it attempts to incorporate an interactive component, the visualization is messy and difficult for users to hover over specific points, which overlaps each other.

Data collection and cleaning

The datasets we have chosen to focus on for this project is available publicly on data.gov.sg.

The datasets came in four separate CSV files:

1. Resale flat prices in Singapore based on its approval dates from January 2000 to February 2012
2. Resale flat prices in Singapore based on its registration dates from March 2012 to December 2014
3. Resale flat prices in Singapore based on its registration dates from January 2015 to December 2016
4. Resale flat prices in Singapore based on its registration dates from January 2017 onwards (up to March 2022).

The four CSV files were joined, and certain variables were augmented to fit our models. We decided to remove flat_model as it was observed in similar works that it is not a key feature when predicting resale price. Table 1 shows the variables present in our dataset, followed by a short description of what it represents and how we derived it.

[5] Chitty, C. (n.d.). 9 things to consider before buying a HDB flat. PropertyGuru. https://www.propertyguru.com.sg/lifestyle/article/3/9-things-to-consider-before-buying-a-hdb-flat

[6] Sengkang Singapore - latest guide and real estate information, places of interest & things to do. (n.d.). Singapore Property and Real Estate for Sale & for Rent | 99.Co. https://www.99.co/singapore/neighbourhoods/sengkang

[7] López, C (2020). Singapore's government will pay people to have children during the pandemic. https://www.insider.com/singapore-will-pay-people-to-have-children-during-the-pandemic-2020-10#:~:text=Singapore's%20government%20is%20offering%20to,sure%20about%20their%20financial%20stability.

[8] Why you should consider living in Sembawang or Sengkang. (2019, June 20). PropertyGuru. https://www.propertyguru.com.sg/property-guides/why-you-should-consider-sembawang-or-sengkang-15822

[9] HDB Map Services. (2019). https://services2.hdb.gov.sg/web/fi10/emap.html

[10] HDB price trends, will housing prices drop or rise in 2020?. (2019). https://www.teoalida.com/singapore/hdbprices/

Table 1: Variables and their descriptions after cleaning the joint dataset

Variable Name	Description
month	The year/month of the transaction.
town	All inputs should only contain "SENGKANG". This was done by filtering the dataset using the feature town.
flat_type	We have decided to focus on only "4 ROOM" and "5 ROOM" room types due to insufficient data for other types.
block	Block number of the flat.
street_name	Street name of the flat.
mean_storey	The values are calculated by taking the mean of the values given in the original feature, storey_range.
floor_area_sqm	Size of the flat measured in square meters.
lease_commence_date	Year the lease started from.
resale_price	Transaction price of the flat.
accom_cpi	The value of the monthly accommodation Consumer Price Index (CPI) in the month of the transaction, with 2019 as base year[11]. These values are to account for inflation and will be used to normalise prices. To calculate adjusted prices, we divided price by their respective accom_cpi value, multiplied by 100.
adj_resale_price	The adjusted resale prices after accounting for inflation.
price_per_sqm	They are calculated by dividing the resale_price by floor_area_sqm. Historically, larger flats will tend to fetch higher prices due to a bigger floor area. To account for the difference in flat sizes in our dataset, we have decided to include this as a factor for a fairer comparison.
adj_price_per_sqm	The adjusted prices per square metres after accounting for inflation.
clean_remaining_lease	This is a simplified version of the original remaining_lease values listed in the dataset. We recalculated the remaining lease by deducting the difference between the year of resale and lease_commence_date from the default 99 years HDB lease.
time_to_mrt	Time taken in minutes to walk from the block to the nearest MRT/LRT station. To calculate this value, we made use of both Google Maps Platform (GMP) *Places API* and *Distance Matrix API*[12].
time_to_hawker	Time taken in minutes to walk to the nearest hawker centre. The way this is calculated is similar to time_to_mrt.
time_to_mall	Time taken in minutes to walk to the nearest shopping mall. The way this is calculated is similar to time_to_mrt.

After taking a closer look into the data, our group removed outliers and filtered out data that we deem unnecessary for model training. To consider the possible fluctuations in resale prices due to the 2008 Great Financial Crisis, we decided to investigate the linear trend from 2007 to 2012.

After analysis, we have decided to remove data from December 2008 to July 2009 as it will give the best linear fit while minimizing the amount of data removed[13].

Models

Our application's main approach is to use regression-based models as listed in the following page.

[11] Srx.com.sg (2022). Singapore Property Heat Map | Singapore Real Estate and Property Exchange. [online] Available at: https://www.srx.com.sg/heat-map?link=heatMap&mapView=true&propertyType=81&district=0&hdbTown=0&keywords=&postal=0&radius=0&projectsTabIndex=0&nearbyTabsIndex=0&pageIndex=0&teamId=&agentId=&locationSearch=true&selectedDistricts=0&selectedHdbTowns=0&selectedMrts=0&maxSale=0&maxRent=0&minSize=0&bedroom=0&propertySubtype=0&sortBy=volume&show=true

[12] Please refer to Annex A for the full details to how this process is done.

[13] Please refer to Annex B for further explanation and graphical representations.

Linear Regression

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.

Random Forest Regression

Random Forest uses an ensemble method to allow for both regression and classification tasks. The ensemble method leverages on a group of decision trees to aggregate the final output instead of using just a singular tree. Each tree is formed using a Bootstrap Aggregation technique, or also known as bagging, where it is trained on a sample that is randomly sampled with replacement from the dataset[14]. Hence, this allows for high accuracy and low training time[15].

Implementations

Linear Regression

Data Preprocessing: For us to use any regression-based model, we have to ensure that our data frame only consists of numerical data. As such, there is a need to reformat two variables: month and flat_type. The variable month was converted into the number of months with reference to the first entry in January 2000 (number of months passed = number of months passed * 12 + month in the data). For flat_type, we can convert it into a column called "flat_4_room", which will contain either 1 or 0. If the selected unit is "4-ROOM", it will be labelled as 1, else it will be labelled as 0. To prevent overfitting due to the number of variables, we only chose to use 8 out of 17 variables for the model training portion. These are namely month, mean_storey, adj_price_per_sqm, clean_remaining_lease, time_to_mrt, time_to_hawker, time_to_mall and flat_4_room. Between the adj_price_per_sqm and price_per_sqm, we chose to predict price_per_sqm, as it is a common metric of measurement for comparison between prices of flats.

Training: Training was performed on Google Colaboratory, in which Python was used to implement the model. We made use of packages from sklearn library. Using the predefined train_test_split() function, we are able randomize and split the dataset into their training and testing datasets. This was done in a 70:30 split, with 70% being used for training.

Results: The linear regression model gave a R^2 score of 0.5853, and an adjusted R^2 score of 0.5851. The Root Mean Square Error (RMSE) between the training data and the model's prediction, as well as between the testing data and the predicted values gives 465.96 and 466.30 respectively. The RMSE tells us the average difference between the predicted and actual values of the model. From the results, we can conclude a prediction accuracy of 58.51%, with an average deviation of $466 across both training and testing dataset. This deviation is about 10% of the actual values, which averages to about $4000, based on 2022 data.

Random Forest Regression

Data Preprocessing: 7 out of 17 features were chosen as input attributes, namely month, flat_type, mean_storey, clean_remaining_lease, time_to_mrt, time_to_hawker and time_to_mall, to predict adj_price_per_sqm. The dataset was split by year. Initial testing of using a 70:30 split resulted in a testing R^2 score of 0.188 while RMSE is at 681.0. After exploring different splits, it was chosen for the testing set to consist of the most recent data obtained this year, whereas the training set consist of previous years. This was done to make sure predictions are relevant to present prices. The final ratio of training to testing is approximately 99:1.

Training: Training was done in Python using the scikit-learn package and optimized on CPU using the sklearnex module. The validation method chosen was the Out-of-Bag (OOB) method, while R^2 score and RMSE were chosen as the evaluation metrics. The advantage of using OOB is that the whole dataset is used for training and error estimation, which makes it more performant than cross-validation type methods, where a subset of the whole dataset is left out of training for error estimation[16]. Hyperparameter tuning was also explored using the Grid Search method, by tuning the number of trees and maximum depth any tree. The tuning was performed using 10-fold cross validation and evaluated using R^2 score for accuracy.

Results: The baseline model gave a training OOB R^2 score of 0.911, with the test set R^2 score dropping to 0.585 and RMSE of 448.7. This means the baseline model has an accuracy of 0.585 and deviates $224.3 from the predicted price, with 87.8% of the testing predictions falling within 10% of the test set. The Grid Search return a value of 'None' for maximum depth and 200 trees as the best parameters. Using those parameters, test set R^2 slips to 0.582, and RMSE of 450.1, with 58.5% of the testing predictions falling within 10% of the test set.

[14] Biau, G. (2012). Analysis of a Random Forests Model. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[15] Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., & Ramachandran, M. (2021). A Comparative Study of Linear, Random Forest and AdaBoost Regressions for Modeling Non-Traditional Machining. Processes, 9(11), 2015. doi: 10.3390/pr9112015

[16] Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. PLOS ONE, 13(8), e0201904. doi: 10.1371/journal.pone.0201904

Results and Applications

Results

The results obtained from both Linear and Random Forest Regression are not good. Putting the accuracy into perspective, the result from each model implies that approximately 1 in 2 predictions will be accurate, which leaves much room for improvement. The RMSE scores obtained from both models reflect that there is approximately a 10% deviation from the predicted and actual values, which could potentially be a huge amount after the predicted value is multiplied by the floor area of the flat.

Applications

Our model can benefit different groups of buyers Singapore, including those looking for new homes and those buying for investment purposes. Additionally, given more data, the model can be used to predict prices for other types of housing, such as condominiums and private houses, as well.

Ethical Implications of Model

Artificial stupidity[17] is an issue of our model. Before training our model, we removed anomalous data caused by unexpected events to ensure accurate model predictions. However, if these events were to happen again, the model will not be able to account for them. This could lead to inaccurate price predictions and false information for users.

Bias is another key concern. Machine learning systems can, intentionally or inadvertently, result in the reproduction of already existing biases[18]. As the dataset was not collected by us, we are unable to confirm if there are presences of biased data which could lead to incorrect predictions. Additionally, there might be instances where our model classifies certain traits with a lower price due to limited data present for it. This might create false hope in users, and this is not ideal.

Limitations and Future Prospects

Limitations

We detected a few variables being dependent on one another. One example is price_per_sqm being dependent on both resale_price and floor_area_sqm. Multicollinearity amongst feature variables is unfavourable. However, taking out price_per_sqm resulted in a negative prediction for resale_price, vice versa. The difference in accuracy is apparent in both all proposed models.

Travel time is calculated by taking the distance between a particular street_name and a selected amenity. Since streets can stretch up to long distances, using street_name as the point of reference could give inaccurate calculations.

Future Prospects

Engaging a data domain expert for a more detailed analysis of the problem will help to improve our models and make better predictions in the future. Additionally, to mitigate the issue of multicollinearity, we could perform other kinds of analysis, such as Principal component analysis[19].

Instead of focusing only on the time taken on foot from street_name to the amenities, we can calculate the travel duration of other modes of transportation, including time taken via car or bus, to account for faster routes. And to obtain a more precise travel time, the full address of the resale flats could be used.

Conclusion

Our model aims to mainly help buyers looking for a suitable place in the resale flat sector. Although our models lacked accuracy in certain aspects, we hope to use this concept to help home buyers feel more confident when buying a house. As Singapore's population size continues to grow, and projected to reach 5.86 million in 2026[20], there will be a rising need in housing as well.

Team Reflections and Roles

Lionel: Starting out this project, I had ambitious goals to create something that could potentially be used to address an issue in Singapore. However, after much discussion and working on the problem formulation, I found that it was quite a difficult task to do, due to the difficulty of narrowing down a problem into a machine learning problem. Through working with my teammates, I realized the importance of having a team with diverse skillsets. I had a teammate who could work with the Google API and another who could quickly manipulate data. There are more projects that I would like to explore in the future, in terms of practical applications, and this project has given me a good start!

Cheng Hui: Before embarking on this project, I usually work with convolution-based networks for my applications. I am

[17] Bossmann, J (2016). Top 9 ethical issues in artificial intelligence. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

[18] Stahl, B. C. (2021). Ethical Issues of AI. SpringerBriefs in Research and Innovation Governance, 35–53. https://doi.org/10.1007/978-3-030-69978-9_4

[19] Rekha, M (2019). MLmuse: Correlation and Collinearity — How they can make or break a model. https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a

[20] Statista. (2021, November 16). Total population of Singapore 2026. https://www.statista.com/statistics/378558/total-population-of-singapore/

grateful to have this opportunity to explore regression-based models and especially work on a topic that is on many young Singaporean adult's mind (housing). Through working with my team, I came to a deeper appreciation of data cleaning to make my job easier when implementing the model, and experience working with Google Maps API to further dig into possible reasons why certain flats might be more desirable. There are definitely more things we can try but I am happy to get to explore other types of ML fields while working with other people with different expertise.

Desmond: Through this project, I was able to get a glimpse of how machine learning can help consumers. I also got to expose myself to critically thinking how our model would benefit someone. I also got to learn new data collection methods such as tapping into the Google Cloud APIs. All in all, it was an eye-opening experience for me knowing machine learning can be applied to almost every domain when designed carefully.

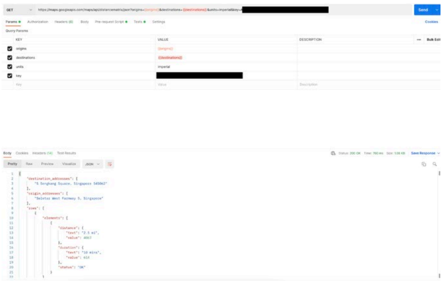
Rachel: In general, I gained a lot of knowledge on the property market during my research on the topic. This is especially so since skewed and anomalous data is abundant in the resale dataset. Without researching about them, it would be difficult to make decisions on how much data should be adjusted or removed. I also got to learn a bit about the model training process during our discussions, which is very new but truly insightful for me.

Wang Chen: Through this project, I got to learn more about how previous research on the topic can help us to better understand the needs of the target audience we are aiming to help. It effectively helps us to choose the variables we would like to focus on training and keep us grounded throughout the whole project.

Member	Roles in project
Lionel	Linear Regression model implementation, report writing
Cheng Hui	Random Forest Regression model implementation, report writing
Desmond	Google Maps API, report writing
Rachel	Data cleaning, report writing
Wang Chen	Data cleaning, report writing

Annexes

Annex A:
Through the Google Maps Platform, we can utilize their APIs to help locate nearby amenities as well as calculate the walking duration to these amenities. We used Postman and imported a Google Maps Platform workspace to aid us in running this API. Once we input our parameters, Postman sends a request to the GMP cloud and returns a response in JSON format containing all the values we need.



https://maps.googleapis.com/maps/api/distancematrix/json?origins={{origins}}&destinations={{destinations}}&units=imperial&key=[redacted]

	KEY	VALUE	DESCRIPTION
☑	origins	{{origins}}	
☑	destinations	{{destinations}}	
☑	units	imperial	
☑	key	[redacted]	
	Key	Value	Description

Status: 200 OK Time: 760 ms Size: 1.06 KB

```
{
    "destination_addresses": [
        "5 Sengkang Square, Singapore 545062"
    ],
    "origin_addresses": [
        "Seletar West Farmway 5, Singapore"
    ],
    "rows": [
        {
            "elements": [
                {
                    "distance": {
                        "text": "2.5 mi",
                        "value": 4067
                    },
                    "duration": {
                        "text": "10 mins",
                        "value": 614
                    },
                    "status": "OK"
                }
            ]
```

Figure 1: Postman interface and JSON output file at the bottom

We first use the Places API to locate the nearby amenities. Subsequently, we stored the coordinates of these amenities into a .csv file. These coordinates then serve as destination input for the second part of this whole data collection process. Here, we use the Distance Matrix API to get the duration it takes to travel from the HDB to the amenity. We then store each of the JSON responses. Using these responses, we took the duration parameter and compiled it into a .csv file. These values are then added to our training model.

Annex B:

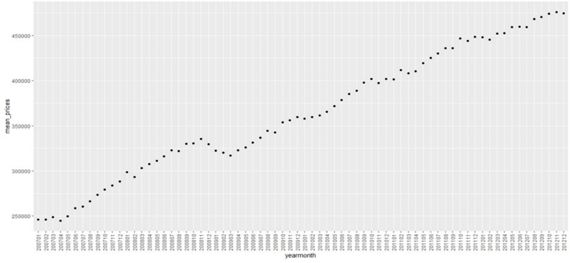


Figure 2: Plot before adjusting for anomalies

According to the summary fit results, $R^2 = 0.952$.

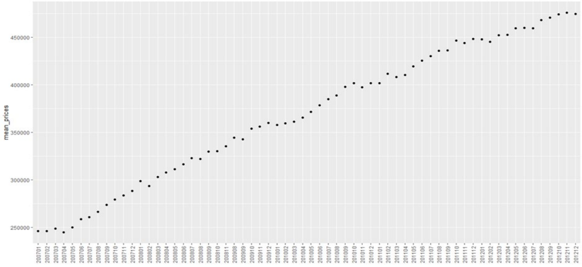


Figure 3: Plot after adjusting for anomalies

After removing all data points from December 2008 to July 2009, the summary fit results show $R^2 = 0.9605$, which is an improved linear trend.