BT 4222

Mining Web Data for Business Insights

Semester II AY2122

Group 23 Final Project Report

Group Members:

Jolene Lee Jia Yi (A0223067L)

Sie Jie Xiang (A0217404N)

Jeyakumar Jegan (A0217456B)

Lim Hao Shen Keith (A0223766Y)

Github link: https://github.com/Keithlimhs/BT4222

Background:

Demand for Build-To-Order (BTO) flats jumped 70 per cent in 2020 with approximately 87,800 applications, while there were 51,400 applications made in 2019 and 38,500 in 2018. However, the number of BTO flats launched did not keep up with demand, with only 16,800 BTO flats launched in 2020[1].

Over 80% of Singapore's residents live in Housing Development Board (HDB) flats, which were designed by the Singapore government in 1960 to resolve the need for urban, comprehensive housing. Because of the scarcity of land in Singapore, HDB flats are generally built as apartment-style units, often in high rise buildings so as to make efficient use of space. Owning a HDB flat for most Singaporeans happens typically as soon as they get married. One unique feature about HDB flats is the 99 year lease associated with it: The lease of the flat is 99 years long, starting from the completion of the flat, with the land reverting back to the government upon expiry. This means that purchasing a HDB flat in Singapore doesn't mean owning the flat in a traditional sense, but rather, leasing the flat from the government for the specified duration of the remaining lease.

Nonetheless, it doesn't mean that Singaporeans have to stay in the owned HDB flat for the rest of one's life. HDB flats can be sold (albeit after 5 years if the flat was purchased brand new), and therein lies the heart of this mini project — the HDB flat resale market. As a young adult perhaps looking to own an HDB flat sometime soon, understanding the prices behind HDB resale flats and predicting them seemed an interesting project to work on.

Problem statement:

With the increasing demand of housing by married couples, coupled with the lack of supply of BTO flats launched, resale flats are becoming a much more viable option in recent years. As such, we would want to build a machine learning model to be able to predict the price of a resale flat given its features such as location, model, size etc, so that married couples would be more informed of the estimated price of a specific resale flat.

Dataset

HDB dataset:

We obtained the HBD dataset from the Singapore government website. The dataset was created in 2015, and updated monthly. It is currently managed by the Housing Development Board (HDB), and the data is based on the date of registration for the resale transactions. There are 5 separate csv files in total: 1990 - 1999, 2000 - Feb 2012, March 2012 - Dec 2014, Jan 2015 - Dec 2016 and Jan 2017 onwards. The initial variable types of the dataset before cleaning are given below.

HDB dataset (before cleaning)

Variable	Variable Type	Description
month	object	Month of transaction
town	object	Planning Area
flat_type	object	Number of rooms

[1] https://www.straitstimes.com/singapore/housing/demand-for-build-to-order-hdb-flats-jumped-in-2020-with-87800-applications

block	object	Block number
street_name	object	Street name
storey_range	object	Range of stories
floor_area_sqm	float64	Floor area per square metre
flat_model	object	Type of flat
lease_commence_date	int64	Lease commencement date
remaining_lease	object	Number of years left on lease
resale_price	float64	Price at which flat was resold

Table 1.1

From the above dataset, we observed that some of the variable types are not in an appropriate type. For example, the month column is an object type, but ideally we would want this column to be in a datetime type. The remaining lease column is also in an object datatype because the values are stored as "XX years XX months". We will have to convert this column to numerical variable type before training the machine learning model.

Data Cleaning

As mentioned above, some of the variable types are not in an appropriate type. Before we go ahead to change the type of the columns, we must first check for duplicates and null values in the entire dataset. Upon inspection, 252 duplicates were found in the dataset. This could be due to human error when the dataset was made. These duplicates were dropped from the dataset, and we found that there were no null values in the dataset as well. Thereafter, we can proceed to clean the data. Firstly, the month column was changed to datetime type, and the remaining_lease column was converted to float by calculating years in decimal form.

Amendments from project proposal:

In our project proposal, we mentioned that we would use all 5 csv files that we obtained from the Singapore government website. However, upon further analysis of the datasets, we decided to only utilise the dataset from Jan 2017 onwards. This is because the dataset from Jan 2017 has 120466 rows, which we feel is sufficient data for model training. Moreover, when we inspected the storey_range column of the resulting dataset joined from the 5 csv files, we noticed that there are storey ranges which overlapped. On inspection of the storey_ranges column only for the dataset of Jan 2017 onwards, we found out that the storey ranges given were of specific categories, for example, “01 TO 03”, “04 TO 06”, “07 TO 09” etc. Hence this gives us the idea of applying Label Encoding to this column as a preprocessing step before training the machine learning model. However, when we joined all 5 csv files into one dataset, we realised that there were overlaps in the categories for storey ranges. For example, there is a category for “01 TO 03” and for “01 TO 05”. This is because the categories for storey ranges are different for the 5 different csv files, and this may hinder the label encoding process applied later on. Hence we decided that we should only make use of the dataset from Jan 2017 onwards.

The second amendment we made from the project proposal was the use of the MRT dataset. We mentioned in our proposal that we would scrape data about the MRT stations in each location which corresponds to the “town” column from the HDB dataset from this wikipedia page. However, upon analysing the MRT data, we realised that there were inaccuracies in the dataset obtained.

14	Geylang	EW7 Eunos EW8 CC9 Paya Lebar EW9 Aljunied CC7...	NaN	8	0	8
15	Hougang	NE13 Kovan NE14 CR8 Hougang CC11 Tai Seng	CR7 Defu CR9 Serangoon North	3	2	5
16	Jurong East	NS1 EW24 JE5 Jurong East EW25 Chinese Garden	JS12 Jurong Pier JE3 Bukit Batok West JE4 Toh...	2	5	7
17	Jurong West	EW26 Lakeside EW27 JS8 Boon Lay EW28 Pioneer	JS5 Corporation JS6 Jurong West JS7 Bahar Jun...	3	5	8
18	Kallang	EW10 Kallang EW11 Lavender NE7 DT12 Little In...	TE23 Tanjong Rhu TE24 Katong Park	9	2	11
19	Lim Chu Kang	NaN	NaN	0	0	0
20	Mandai	NaN	NaN	0	0	0
21	Marina East	NaN	TE22A Founders' Memorial	0	1	1
22	Marina South	NaN	TE21 Marina South TE22 Gardens by the Bay	0	2	2
23	Marine Parade	NaN	TE24 Katong Park TE25 Tanjong Katong TE26 Mar...	0	4	4
24	Museum	NS24 NE6 CC1 Dhoby Ghaut CC2 Bras Basah DT20 ...	NaN	4	0	4
25	Newton	NS21 DT11 Newton	NaN	1	0	1
26	North-Eastern Islands	NaN	NaN	0	0	0
27	Novena	NS20 Novena DT10 TE11 Stevens	CC18 Bukit Brown[lower-alpha 2] TE10 Mount Pl...	2	2	4
28	Orchard	NS22 TE14 Orchard NS23 Somerset	TE13 Orchard Boulevard	2	1	3
29	Outram	EW16 NE3 TE17 Outram Park NE4 DT19 Chinatown ...	TE18 Maxwell	3	1	4
30	Pasir Ris	EW1 CR5 CP1 Pasir Ris	CR3 Loyang CR4 Pasir Ris East CP2 Elias	1	3	4
31	Paya Lebar	NaN	NaN	0	0	0
32	Pioneer	EW29 Joo Koon EW30 Gul Circle	NaN	2	0	2
33	Punggol	NE17 CP4 PTC Punggol	NE18 Punggol Coast CP3 PE4 Riviera	1	2	3
34	Queenstown	EW19 Queenstown EW20 Commonwealth EW21 CC22 B...	NaN	9	0	9
35	River Valley	NS22 TE14 Orchard	TE13 Orchard Boulevard TE15 Great World	1	2	3
36	Rochor	EW12 DT14 Bugis NE7 DT12 Little India NE8 Far...	NaN	6	0	6

As can be seen here, Paya Lebar was indicated as having no MRT stations in the area. This is obviously incorrect since we know that there is an MRT station there. Furthermore, some areas were indicated as having multiple MRT stations in that area, but the page does not explain to us how the number of MRT stations for each planning area was obtained. Hence we decided not to use this MRT dataset anymore.

Feature Engineering

The notable features of the HDB dataset is given above in Table 1.1. One key point to note is that these features are intrinsic in nature, meaning that these features describe the HDB unit itself, such as its location, size etc. However, the resale price of a HDB unit is dependent on certain external factors as well. In hypothesising factors that could affect the price of HDB resale flats, our team brainstormed 5 prominent factors:

1. Distance to the nearest MRT

This is an indicator of how well-connected the HDB unit is to the MRT network in Singapore, which serves as the main form of public transport for Singaporeans. We feel that the distance to the nearest MRT would affect the resale price of a HDB unit as it shows how convenient and accessible the unit is.

2. Distance to the nearest mall

Since shopping malls in Singapore provide many services and products such as food, shopping, groceries, entertainment etc, we feel that the distance to the nearest shopping mall would affect the resale price of the HDB unit, since it shows how convenient the HDB unit is to nearby amenities and services.

3. Distance to the nearest primary school

The distance to the nearest primary school may affect the resale price of a HDB unit as well because having a primary school nearby could be a great convenience for parents as their young children have the option of walking home after school instead of having alternative transport arrangements.

4. Distance to Central Business District (CBD)

The distance to the CBD area may affect the resale price of a HDB unit as well, since the CBD area houses many offices for Singapore's workforce, hence it would be an indicator of how convenient parents can go to work, if they work in the CBD area.

5. Whether the estate is matured or not
An indicator on whether the planning area is a matured estate may affect the resale price of a HDB unit as well, because Singaporeans view matured estates as better due to greater proximity to amenities[2].

Data Collection

Before we can dive straight into the feature engineering that we have identified above, we must first gather the required data that we need. In order to calculate distances of the HDB unit to the nearest MRT, nearest Primary School, and nearest mall, we must first get the coordinates of each HDB unit in our HDB dataset. Our team decided to utilise the open source and free OneMap API that was created by the government. In order to use the API to retrieve the latitude and longitude coordinates of each HDB unit, the API call requires a query string which takes as input a parameter "searchVal", which is an address to be searched by the API. "returnGeom" and "getAddrDetails" were set as Y in order for us to retrieve the coordinates.

Parameters

Variables	Description
`searchVal` *Required*	Keywords entered by user that is used to filter out the results.
`returnGeom {Y/N}` *Required*	Checks if user wants to return the geometry.
`getAddrDetails {Y/N}` *Required*	Checks if user wants to return address details for a point.
`pageNum` *Optional*	Specifies the page to retrieve your search results from.

In order to obtain the address needed for the "searchVal" parameter, we created a new column, "address", in the HDB dataset, which consists of the "block" column joined with "street_name" column.
Moreover, since there are multiple transactions in the same block in a HDB as there can be multiple units sold within a block, we get the unique addresses and iterate through the list of unique addresses, using each address as input for "searchVal" parameter. Thereafter, we can obtain the latitude and longitude coordinates from the resulting json returned from the OneMap API calls. Hence we have successfully obtained the latitude and longitude coordinates of each unique address in the HDB dataset.

[2] https://www.99.co/singapore/insider/mature-estate-vs-non-mature-estate/

```
{'found': 1,
 'totalNumPages': 1,
 'pageNum': 1,
 'results': [{'SEARCHVAL': '429 PASIR RIS DRIVE 6 SINGAPORE 510429',
   'BLK_NO': '429',
   'ROAD_NAME': 'PASIR RIS DRIVE 6',
   'BUILDING': 'NIL',
   'ADDRESS': '429 PASIR RIS DRIVE 6 SINGAPORE 510429',
   'POSTAL': '510429',
   'X': '41734.3036202751',
   'Y': '39121.6563188341',
   'LATITUDE': '1.37007368495954',
   'LONGITUDE': '103.956730877173',
   'LONGTITUDE': '103.956730877173'}]}
```


JSON returned from OneMap API

After obtaining the latitude and longitude coordinates of each HDB address, we have to gather the coordinates of all MRT, Primary schools, and shopping malls in Singapore. Using the BeautifulSoup library, we scraped for the lists of all MRT stations, Primary Schools, and shopping malls online. Similar to the HDB addresses, we iterated through the lists of MRT stations, Primary Schools, and shopping malls, and input the MRT/primary school/shopping mall name into the “searchVal” parameter of the API call, thereby obtaining the latitude and longitude coordinates of each MRT/Primary school/shopping mall.

With the latitude and longitude coordinates of all unique HDB addresses, MRT Stations, Primary Schools, and Shopping malls, we can then calculate and find the distance to the nearest MRT station, Primary School, and shopping mall from each unique HDB address.

For each unique HDB address, we loop through the coordinates of all the MRT stations that we have, calculating the distance of the HDB address to each of the MRT stations in Singapore. The minimum distance and the MRT station that corresponds to the minimum distance is then returned. The same is repeated for the list of Primary schools and shopping malls. The distances were calculated using the Haversine formula from the Haversine package, which calculates the distance between 2 points on a curved surface, in kilometres.

Finally, the distance to CBD was calculated as well using these coordinates = (1.287953, 103.851784), which was returned from a google search of the coordinates of Downtown Core of Singapore.

The dataset of each unique HDB address with the corresponding nearest MRT, Primary school, and shopping mall, and their associated distances is shown below.

	address	latitude	longitude	nearest mrt	dist to nearest mrt	nearest school	dist to nearest school	nearest mall	dist to nearest mall	dist to cbd
0	406 ANG MO KIO AVE 10	1.362005	103.853880	Ang Mo Kio MRT Station	1.003997	Townsville Primary School	0.185706	AMK Hub	1.003115	8.237463
1	108 ANG MO KIO AVE 4	1.370966	103.838202	Ang Mo Kio MRT Station	1.267607	Ang Mo Kio Primary School	0.228347	Broadway Plaza	0.870651	9.353344
2	602 ANG MO KIO AVE 5	1.380709	103.835368	Yio Chu Kang MRT Station	1.071179	Mayflower Primary School	0.517733	Broadway Plaza	1.530164	10.474184
3	465 ANG MO KIO AVE 10	1.366201	103.857201	Ang Mo Kio MRT Station	0.945529	Teck Ghee Primary School	0.694710	myVillage At Serangoon Garden	0.879263	8.721611
4	601 ANG MO KIO AVE 5	1.381041	103.835132	Yio Chu Kang MRT Station	1.094010	Mayflower Primary School	0.557204	Broadway Plaza	1.574032	10.515191
...	...	...	...	...	...	...	...	...	...	...
9194	110A DEPOT RD	1.281428	103.809243	Telok Blangah MRT Station	1.193602	Blangah Rise Primary School	0.592914	Alexandra Central	0.805653	4.784510
9195	49 JLN BAHAGIA	1.327660	103.856373	Boon Keng MRT Station	1.065611	Hong Wen School	0.678932	HDB Hub	0.875751	4.444582
9196	121A CANBERRA ST	1.447602	103.833518	Canberra MRT Station	0.658068	Chongfu School	1.189041	Canberra Plaza	0.587766	17.867919
9197	844 TAMPINES ST 82	1.352103	103.936534	Tampines MRT Station	0.762443	St. Hilda's Primary School	0.342722	Our Tampines Hub	0.492275	11.817074
9198	502C YISHUN ST 51	1.417323	103.841554	Khatib MRT Station	0.953151	Naval Base Primary School	0.348981	Wisteria Mall	0.083069	14.430197

9199 rows × 10 columns

Dataset containing coordinates and distances of each unique HDB address

This dataset is then merged with the HDB dataset on the “address” column. Lastly, we engineered the column “matured”, which indicates if the HDB is in a matured estate. The list of matured estates were obtained here, and each row was indicated as 1 in “matured” if the “town” observation is in the list of matured estates.

The data schema of the final dataset is shown below

Variable	Variable Type	Description
month	datetime	Month of transaction
town	object	Planning Area
flat_type	object	Number of rooms
block	object	Block number
street_name	object	Street name
storey_range	object	Range of stories
floor_area_sqm	float64	Floor area per square metre
flat_model	object	Type of flat
lease_commence_date	int64	Lease commencement date
remaining_lease	float64	Number of years left on lease
resale_price	float64	Price at which flat was resold
address	object	Address of HDB unit
latitude	float64	Latitude coordinate

longitude	float64	Longitude coordinate
nearest mrt	object	Nearest mrt from the HDB
dist to nearest mrt	float64	Distance of nearest HDB
nearest school	object	Nearest primary school from HDB
Dist to nearest school	float64	Distance of nearest primary school
Nearest mall	object	Nearest mall from HDB
Dist to nearest mall	float64	Distance of nearest mall
Dist to cbd	float64	Distance to CBD area from HDB
matured	int64	Indicates if the HDB is in a matured estate or not

Table 1.2: Final dataset

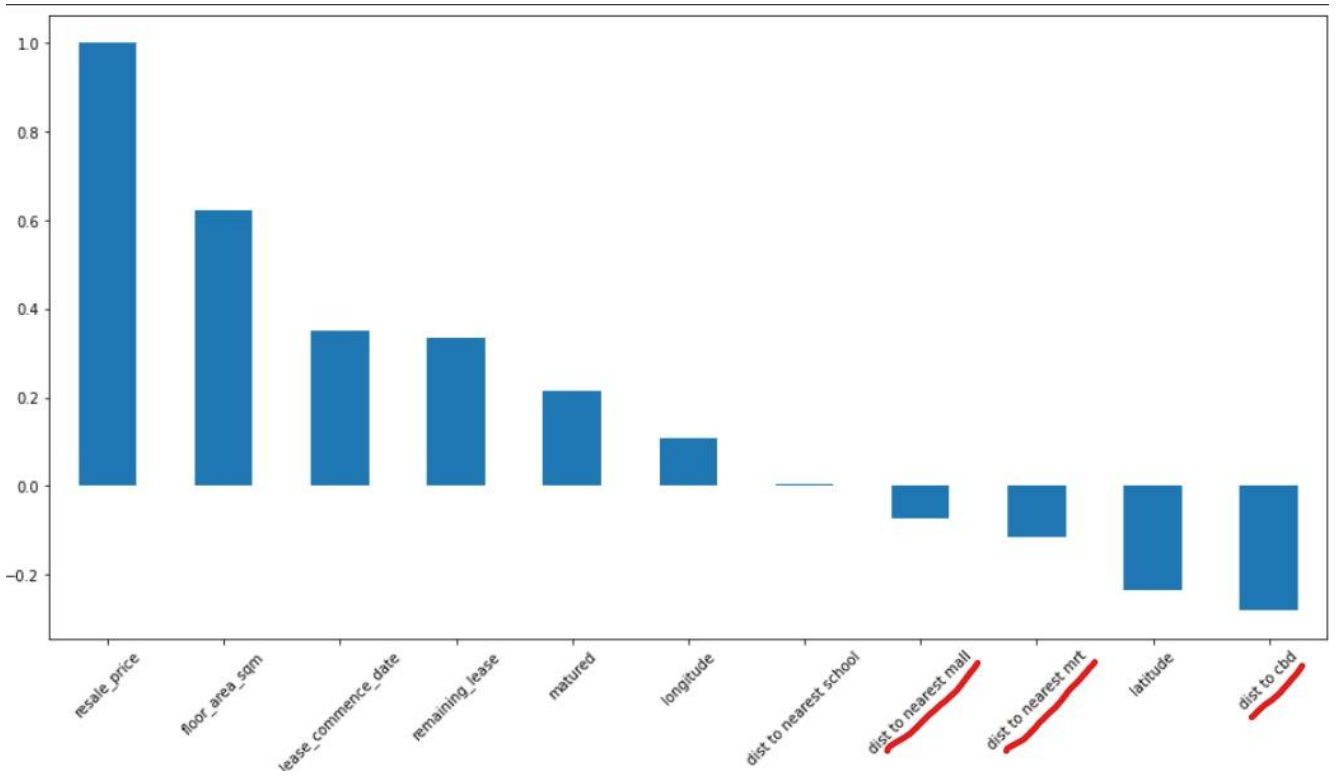
Exploratory Data Analysis (EDA)

Hypothesis 1: Housing location with close proximity to MRT stations might face an increase in price due to convenience that it provides

By plotting a scatter plot of resale price against distance to nearest MRT, we can see that there is a negative relationship between distance to nearest MRT and resale price. Hence HDB units with close proximity to MRT stations can demand a higher resale price.

Hypothesis 2: Houses with close proximity to town areas might have a higher demand and hence higher prices

By plotting a scatter plot of resale price against distance to CBD, we can see that there is a negative relationship between distance to CBD and resale price. Hence HDB units which are closer to the CBD area demand a higher resale price.



Bar plot of correlation between resale price and features

This bar chart summarises the correlation between the features and resale price. As can be observed, the distance to the nearest mall, MRT, and CBD area are negatively correlated with resale price.

Data preprocessing

We applied label encoding to the flat_type and storey_range columns as these columns contain categorical values that are ordinal. A depiction of how the flat_type is label encoded is shown below.

flat type	flat type coded
Multi generation	6
executive	5
5 room	4
4 room	3
3 room	2
2 room	1
1 room	0

Label encoding for flat type

Thereafter, we ran a correlation matrix of the numerical columns and plotted it on a heatmap to visually see each predictor's relationship with one another. From the heatmap, we can see that remaining_lease and lease_commencement_date have a high correlation of 0.99 with each other, and this makes sense since remaining_lease is derived from the lease_commencement_date. Floor_area_sqm has a high correlation of 0.95 with the label encoded flat type, and this makes sense because the higher the label encoded number for flat_type, the bigger the floor area. Due to high correlation, we decided not to use the label encoded flat type and lease_commencement_date as features for machine learning model training. Furthermore, "town" and "flat_model" columns were one-hot encoded as these categorical variables are not ordered. Finally, the set of features used for model training are "floor_area_sqm", "remaining_lease", "dist to nearest mrt", "dist to nearest school", "dist to nearest mall", "dist to cbs", "matured", "storey_range coded", as well as the one-hot encoded "town" and "flat_model" columns.

Machine Learning Models

Since we are trying to predict the resale price of HDB units, our dependent variable would be the "resale_price" column of the HDB dataset. Therefore, this is a one-dimensional regression problem as we are predicting a continuous value, hence we will be taking a supervised learning approach. Furthermore, train_test_split was not used to randomly assign the train and test data because the dataset followed a chronological order, and it made more sense to use past data to predict future data. Therefore, instead of using KFold cross validation, we made use of Time Series Split instead, setting the number of splits to be 3.

Individual Models

We first utilised individual models to set our baseline. We decided to use Support Vector Regression, Decision Tree and Linear Regression as our baseline models. The metrics that we used throughout are the R-squared value and the Root Mean Squared Error (RMSE). Our results show that the linear regression model gave the best results with the highest R-squared and lowest RMSE. The R-squared shows that approximately 80.9% of the variance of our data can be explained by the model.

	Model	R-squared	RMSE
0	SVR	-0.251752	167305.283817
1	Decision Tree	0.687748	83560.944972
2	Linear Regression	0.80924	65312.201676

Ensemble Models

We decided to use 3 ensemble models, 2 of which are Stacked models. Both the Stacked models use Support Vector Regression, Decision Tree, and Linear Regression as the layer 0. The difference between the 2 Stacked models is the meta learner, in which the first Stacked model uses Linear Regression as meta learner while the second Stacked model uses Decision Tree as the meta learner. The last ensemble model we used is Random Forest. The idea of using Stacked models is to compare the performance of the stacked models when heterogeneous weak learners are used together rather than alone.

Unsurprisingly, the ensemble models gave better scores than the individual models, with the exception of the stacked model which uses the decision tree as the meta learner. Random Forest gave the best R-squared and lowest RMSE score among the ensemble models used. The R-squared shows that approximately 87.8% of the variance of our data can be explained by the model.

	Model	R-squared	RMSE
1	Decision Tree as meta learner	0.793683	71402.716636
0	Linear regression as meta learner	0.873331	55947.690612
2	Random Forest	0.878006	54905.654566

Gradient Boosting Models

We used 3 Gradient Boosting models in total, XGBRegressor from the xgboost library, GradientBoostingRegressor from the scikit-learn library, and LGBMRegressor from the lightgm library. The R-squared and RMSE results were calculated after training the models with Time Series Split of the data. Light Gradient boosting machine, LGBMRegressor, turns out to be the best gradient boosting model with highest R-squared and lowest RMSE. The R-squared shows that approximately 88.4% of the variance in the data can be explained by the model.

	Model	R-squared	RMSE
1	Gradient Boosting	0.857353	59371.647954
0	XGBoost	0.858954	59037.44644
2	Light Gradient Boosting	0.884807	53353.226908

Hyperparameter tuning

Since the Light Gradient Boosting model gave us the best score, we went on to perform hyperparameter tuning of the Light Gradient Boosting model using GridSearch cross validation. The search grid that we defined contains the parameters “num_leaves”, with values ranging from 20 to 100 at intervals of 10, “max_depth” with values ranging from 3 to 10 at intervals of 1, and “min_data_in_leaf” with values ranging from 4 to 10 at intervals of 1. We used the R-squared scoring for the Grid Search, as well as Time Series Split for the cross-validation.

Conclusion

The best model with the best set of parameters obtained an R-squared score of 0.889, which is only slightly better than the default values. Using this model, we can find the feature importances of the features used in training the model. The top 5 important features were given as “remaining_lease”, “floor_area_sqm”, “dist to cbd”, “dist to nearest mrt”, and “dist to nearest mall”. Hence these 5 features serve as the strongest predictors of HDB resale prices.