# IT5006:Fundamentals of Data Analytics Project Report

## Group25

## Members:

| Name | ID |
|---|---|
| Xiaodong Yan | A0111696A |
| Aaron Lok Hin Chan | A0117352M |
| Tomohisa Kataoka | A0243462J |

## Table of Contents:



## Link: Kaggle Notebook

https://www.kaggle.com/code/it5006group25/group25

Shared with following usernames in Kaggle:
- AshishDeepak
- A0116612R_ZI-YU

# PART 1: Prediction Report

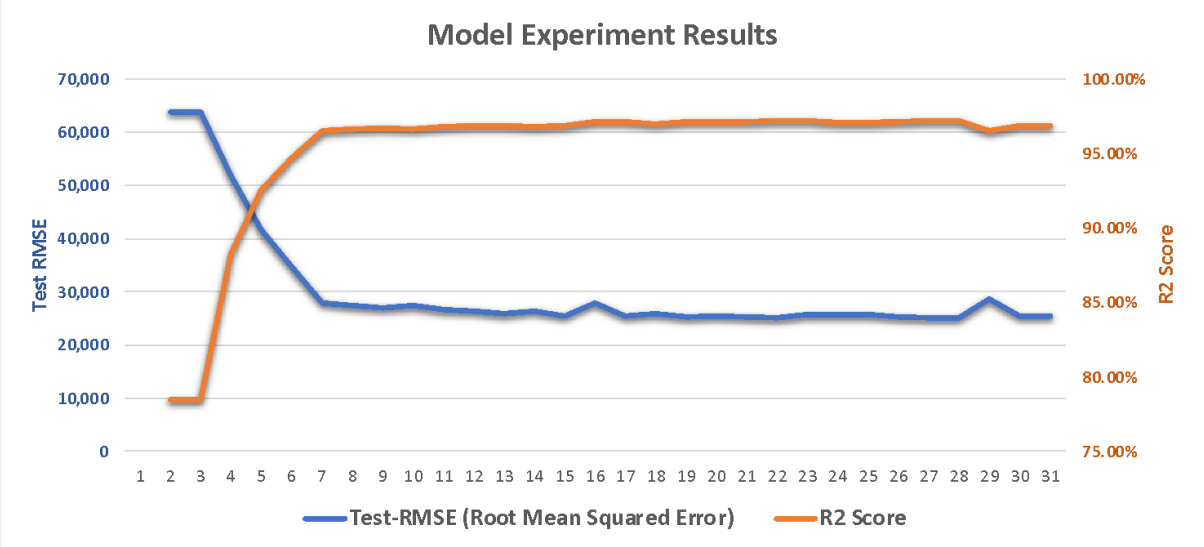
## 1. Data Pre-processing

Response variable: resale_price
Explanatory variables:

| Feature Name | Feature Creation | Type | Processing | |
|---|---|---|---|---|
| | | | Step1 | Step2 |
| town | - | Categorical | One Hot Encoding | - |
| flat_model | - | | | - |
| year | Extracted from sales date | | Label Encoding | Standard Scaled |
| month | | | | |
| quarter | Created from month | | | |
| category_flat_type | Binned to 4 categories | | | |
| floor_area_sqm | - | Numerical | - | |
| avg_storey_range | e.g) 04 to 06 => 05 | | - | |
| remaining_lease | Generated from “date” and “lease_commence_date” | | - | |

## 2. Prediction Methodology

The Models we have tried were **Linear Regression, LASSO regression, Ridge Regression, DecisionTree, Random Forest, XGBoost.** We evaluated these models by mainly **RMSE** (root mean squared error). At first, we just went through each model without hyperparameter tuning and compared the prediction results. After selecting the one that provided best result, we conducted hyperparameter tuning methods (**GridSearchCV, RandomizedSearchCV**) to improve the accuracy of our prediction.




## 3. Results

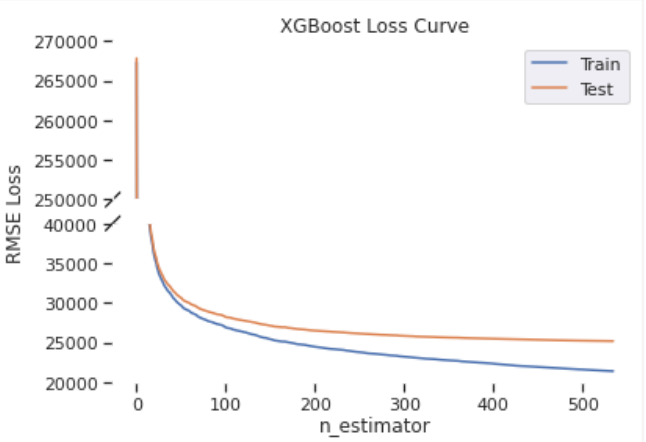
By comparing the results of each model as shown in the Table below, we could get the best performance from XGBoost. As the loss curve shown in the figure, we can not observe

overfitting after hyperparameter tuning. Thus, we are able to expect an appropriate prediction result from this prediction model.

**Table: Model Selection**

| Rank | Model | RMSE |
|---|---|---|
| 1 | XGBoost | 25388.2 |
| 2 | Random Forest | 27393.8 |
| 3 | Decision Tree | 36107.0 |
| 4 | Linear Regression | 63815.3 |
| 5 | Lasso Regression | 63816.7 |
| 6 | Ridge Regression | 63816.7 |

**Figure: XGBoost Loss Curve**




**Our final Kaggle score: 0.97248**

## PART 2: Analytics Report

***"Impact of Downtown Line Opening on Prices of HDB Flats"***

**1. Problem Statement**

To determine the effect of a new MRT line opening in Singapore on the resale value of HDB flats in surrounding areas. As the proximity to MRT stations is a sizable factor in HDB resale prices, we want to further explore the relationship between the two variables (resale price vs MRT proximity). The insights from this analysis could help:

- Buyers and sellers on assessing HDB resale prices, and convenience factors
- Determine the tangible value of being situated near a MRT station

**2. Data**

Based on our proposed problem, we would need to match each HDB unit with its closest MRT station at the time of sale. To do so, we would need:

- Coordinates of each HDB unit (or block, since all units in the same block would share the same coordinates);
- Coordinates of all MRT stations;
- Opening dates of all MRT stations

Opening dates of the MRT stations are required as we need to determine if a MRT station, which is currently present, was available when the flat was sold in the past. If the MRT station did not exist, then we should not account for the presence of that MRT station. A summary of the data sources and data items is provided in the table below.

Table 1: Data Sources

| **Data Required** | **Source of Data** | **Details** |
|---|---|---|
| Coordinates of each HDB Unit | OneMap API | Repeatedly call the OneMap API with each block's address to get the coordinates and postal code as a JSON object. |

| | | |
|---|---|---|
| Coordinates of all MRT Stations | Kaggle (Link) | Readily available data, last updated as of Oct 2021. |
| Opening Dates of all MRT Stations | Wikipedia | Scrapped the opening dates from Wikipedia. |

**3. Methodology**

This study will adopt the use of a difference-in-difference (DiD) regression analysis, where flat transactions will be categorised into four sub-groups: treatment vs. control, before opening vs. after opening of the line, as shown in Table 2. This use of a DiD regression would be similar to that adopted by Diao, Mi, et al. (2017), whereby they had adopted this method to analyse the effects of the Circle Line MRT in Singapore.

Table 2: Treatment vs. Control Groups

| | **Before Opening of DTL** | **After Opening of DTL** |
|---|---|---|
| **Treatment (<=500m of DTL)** | $Group_{Treatment,\ Before}$ | $Group_{Treatment,\ After}$ |
| **Control (>500m from DTL)** | $Group_{Control,\ Before}$ | $Group_{Control,\ After}$ |

The study will focus specifically on the Downtown MRT line (DTL), where the size of the impact would determined by the coefficients for the following: ($Group_{Treatment,\ After}$ - $Group_{Treatment,\ Before}$) - ($Group_{Control,\ After}$ - $Group_{Control,\ Before}$). While the opening of the DTL took place across three dates - 22 Dec 2013, 27 Dec 2015, 21 Oct 2017 (Wikipedia), the period from 2012 to 2018 did not coincide with the opening of any other MRT lines. Hence, the effects estimated in this study can be attributed to the opening of the DTL, and would not be confounded with the opening of any other lines.

The form of the equation is as follows:

$$PSM = B_0 + B_1 TimePeriod + B_2 Treatment + B_3 TimePeriod * Treatment + B_4 OtherVariables$$

B3 is the coefficient of interest, as it would represent the impact of the opening of the DTL. TimePeriod and Treatment are binary variables that equal 1 if the transaction took place after the opening of the MRT station, and if the transaction took place at a flat within 500m of a MRT station respectively.

**4. Results**

The main results are summarised in Table 3 below. For brevity, results for other variables have been omitted.

Two models were used. The difference between Model 1 and Model 2 is the inclusion of a $RemainingLease^2$ variable. The inclusion of this model improved the $Adjusted - R^2$ from 0.7613 to 0.7646. More importantly, the $TimePeriod$ variable, which was statistically insignificant at the 5% level of significance, became statistically significant under Model 2.

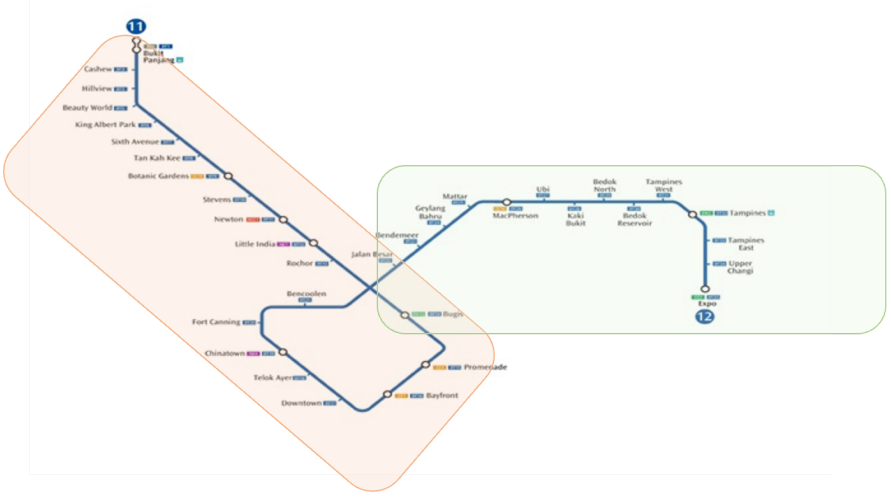
Table 3: Regression Results

| | **Model 1** | **Model 2** |
|---|---|---|
| Adjusted R_Squared | 0.7613 | 0.7646 |

| | | |
|---|---|---|
| Intercept | 2654.4998***<br>(118.9887) | 6605.6962***<br>(180.2534) |
| Remaining Lease | 54.6900***<br>(0.3213) | -54.7927***<br>(3.7851) |
| Remaining Lease_Squared | | 0.7142***<br>(0.0246) |
| Treatment | -228.2454***<br>(15.7654) | -214.7709***<br>(15.6637) |
| TimePeriod | -6.1537<br>(4.6541) | -15.4382***<br>(4.6331) |
| Treatment x TimePeriod | 90.3185***<br>(20.6795) | 79.2672***<br>(20.5405) |

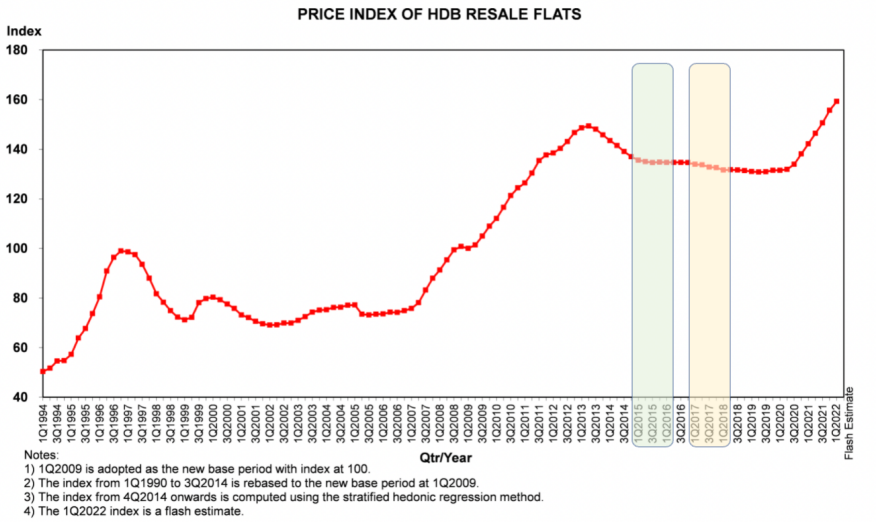
Based on the results of Model 2 in Table 3, flats transacted along the DTL commanded a PSM of $214 lower than flats more than 500m away from the DTL. Based on a check of the DTL stations, it can be noted that most of the DTL stations are either located in expensive areas without HDB flats (orange shaded zone), or industrial/inexpensive areas with HDB flats (green shaded zone), as shown in Figure 1. As a result, it would be reasonable for the coefficient to be negative.

Figure 1: DTL




Before and after the two time periods of Dec 2015 and Oct 2017, resale PSMs have on average fallen by $15, as per the coefficient for $TimePeriod$. Comparing this against the official Resale Price Index (Figure 2) published by the Housing Development Board (HDB), which is the government agency that oversees the public housing market in Singapore, this negative coefficient would also be justifiable. As observed by the green (2015) and yellow (2017) shaded areas, the prices have either fallen or remained constant.

Figure 2: Resale Price Index (HDB)




Finally, the interaction variable $Treatment * TimePeriod$ is the variable of interest for this study. The coefficient of 79.2672 is statistically significant at the 5% level of significance, and indicates that after the DTL had opened, resale flat PSMs have increased, on average, by $79 (rounded), ceteris paribus. Figure 3 shows how this value had been derived through the difference-in-difference methodology.

Figure 3: Interpretation of Results

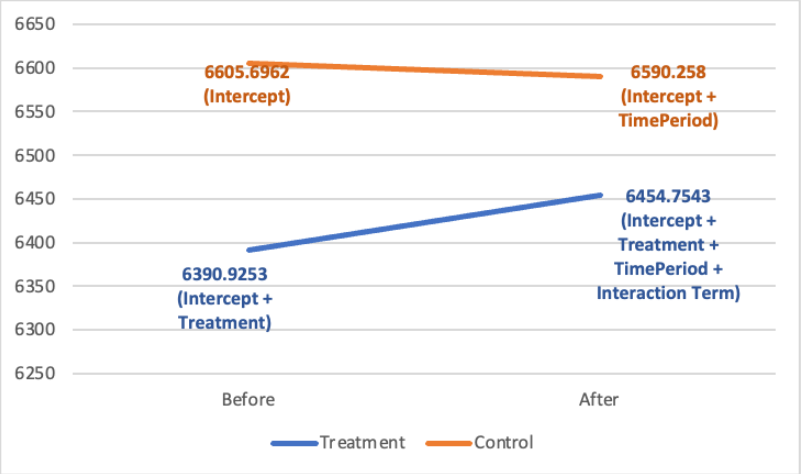



Figure 3 above shows an approach to interpret the results. Before the DTL was built, flats near DTL transacted at a relative PSM of $6,390[1], while flats in other locations transacted at around $214 higher, at $6,605. After the DTL was opened, flats further away from the DTL, which were not affected by the opening, fell by $15 in PSM possibly due to other exogenous factors, to a PSM of $6,590. However, despite the general fall in PSM, flats near the DTL increased in PSM over the same period by $79, which could be attributed to the building of the DTL.

## 5. Robustness

In building the models above, the study had used a threshold distance of 500m to classify flats as being considered to be near the DTL. In Diao, Mi, et al.'s 2017 study, they had used a threshold distance of 400m. To test for the robustness of the results, the models were

[1] The PSMs may seem very different from the typical PSMs of $3,000 to $4,000 as these are intercepts after removing effects of other variables such as location, remaining lease, which may depress the prices respectively.

replicated for threshold distances of 400m and 1,000m. A concise version of the results are summarised in Table 4 below.

Table 4: Robustness Checks

| | 500m (Original) | 400m | 1,000m |
|---|---|---|---|
| Adjusted R_Squared | 0.7646 | 0.7640 | 0.7670 |
| Intercept | 6605.6962***<br>(180.2534) | 6640.5661***<br>(180.470) | 6515.7445***<br>(179.369) |
| Treatment | -214.7709***<br>(15.6637) | -148.4850***<br>(18.650) | -320.7570***<br>(12.424) |
| TimePeriod | -15.4382***<br>(4.6331) | -14.1193***<br>(4.601) | -15.9020***<br>(4.761) |
| Treatment x TimePeriod | 79.2672***<br>(20.5405) | 77.9901***<br>(24.916) | 45.0120***<br>(14.213) |

As shown in Table 4, after adjusting the threshold distance to 400m and 1,000m respectively, the results are still significant and have the same signs, indicating that PSMs did increase for flats nearer to the DTL after the DTL opened. The magnitude of the coefficient for 1,000m is reasonably lower, due to the diluted impact of the DTL on flats further away (i.e. between 500m to 1,000m).

## 6. Conclusion

Based on the analysis, the opening of the Downtown Line in 2015 and 2017 could have resulted in an increase of $79 PSM for HDB flats located within 500m of the new station. Any changes in the prices of HDB flats across Singapore had also been accounted for, hence we could conclude that this estimated increase could be solely attributed to the opening of the Line.

As an extension, we could also analyse the announcement effect, as prices could already have increased after the announcement of the Line and before the opening of the Line. Hence, the true increase might be even higher.

## Works Cited


Diao, Mi, et al. “A New Mass Rapid Transit (MRT) Line Construction and Housing Wealth: Evidence from the Circle Line.” *Journal of Infrastructure, Policy and Development*, vol. 1, no. 1, 2017, pp. 65-89, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2899470.

Housing & Development Board. “Resale Statistics.” *HDB*, HDB, https://www.hdb.gov.sg/residential/buying-a-flat/resale/getting-started/resale-statistics. Accessed 9 April 2022.

Wikipedia. “Downtown MRT line.” *Wikipedia*, https://en.wikipedia.org/wiki/Downtown_MRT_line. Accessed 8 April 2022.