Semester 2, AY2022/23
CS5228

Final Report
By Group 10
Kaggle team name: HelloWorld

Name	NETID
Boran Yang	E1101548
Yankai Fan	E1101541
Weiyi Wu	E1101804
Feiyang Huang	E1101718

Table of Contents

Abstract

In this report, we will introduce how we preprocessed the raw data, conducted exploratory data analysis, tried various models, and evaluated and interpreted model performances for a Singaporean housing resale price competition on Kaggle. The competition contains a main dataset and seven auxiliary datasets. Our goal is to predict the resale price of houses using features such as address, floor area, flat model, and surrounding facilities information provided in the auxiliary datasets. By comparing multiple models with hyperparameters tuned and conducting cross-validation, the experiment demonstrated that CatBoost performed the best among all the models. This project of predicting resale house accurately can allow Singaporeans to make more informed decisions about resale house prices during their purchases and better understand whether a property is overpriced or underpriced. Our best performing prediction model ranked top 3 in the Kaggle competition.

1. Introduction

The ability to accurately predict resale housing prices in Singapore is of significant interest to a variety of stakeholders, including home buyers, home sellers, real estate agents, and policy makers. There are several reasons why research on predicting Singapore resale housing prices is necessary. Firstly, housing is a major asset for many Singaporeans, and resale housing prices may have a significant impact on their financial well-being. Accurate predictions of resale housing prices can help buyers and sellers make more informed decisions. Secondly, the government plays a significant role in the housing market. Accurate predictions of resale housing prices can inform policy decisions related to affordability of public housing. Finally, predicting resale housing prices can also inform the work of real estate agents and property developers. They can make more informed decisions regarding the advertising, sale, and development of properties based on pricing information.

2. Exploratory Data Analysis & Preprocessing

2.1. Overview of Datasets

2.1.1. Main Dataset

The original main dataset contains 431732 records (rows), and each record represents a Singaporean housing transaction. There are 16 different features (attributes) and 1 target label in the main dataset. The summary of each column can be found in the Table 1 below.

Column Name	Data Type	Description
month	Categorical	The year and month in which the transaction occurred. The range is from 2000-01 to 2020-11.
town	Categorical	The town where the transaction property is located.
flat_type	Categorical	The flat type of the transaction property.
block	Categorical	The block of the transaction property.
street_name	Categorical	The name of the street where the house is located.
storey_range	Categorical	The floor level of the housing unit.
floor_area_sqm	Numerical	The floor area of the house.
flat_model	Categorical	The flat type of the transaction property.
eco_category	Categorical	All values are “uncategorized”, it is meaningless.

lease_commence_date	Numerical	The lease commencement date of the property.
latitude	Numerical	The latitude of the location of the house.
longitude	Numerical	The longitude of the location of the house.
elevation	Numerical	All values are “0.0”, it is meaningless.
subzone	Categorical	The subzone where the transaction property is located.
planning_area	Categorical	The planning area where the transaction property is located.
region	Categorical	The region where the transaction property is located.
resale_price	Numerical	The target label. The re-selling price of the property.

Table 1: Summary of Columns of the Main Dataset

As ***storey_range*** contains the range of floor levels and there are similar categories such as ’01 to 03’ and ‘01 to 05’, we took the mean of the range as the new numeric ***storey_range*** variable to be used in the later parts.

2.1.2. Auxiliary Dataset: Commercial Centers

The auxiliary dataset about the commercial centers provides the ***name***, ***type***, ***latitude***, and ***longitude*** of 38 different commercial centers. The ***name*** and ***type*** are categorical attributes, and the ***latitude*** and ***longitude*** are numerical attributes. For the attribute ***type***, there are five different categories.

2.1.3. Auxiliary Dataset: Shopping Malls

The auxiliary dataset about the shopping malls provides the ***name***, ***latitude***, ***longitude***, and ***wikipedia_link*** of 174 unique shopping malls. The ***name*** and ***wikipedia_link*** are categorical attributes, and the ***latitude*** and ***longitude*** are numerical attributes. There are 72 shopping malls missing the ***wikipedia_link***, and it is difficult to conduct any automated analysis on the attribute, so the attribute is not used when we perform integration in Section 2.2.

2.1.4. Auxiliary Dataset: Primary Schools

The auxiliary dataset about the primary schools provides the ***name***, ***latitude***, and ***longitude*** of 191 unique primary schools. The ***name*** is categorical attribute, and the ***latitude*** and ***longitude*** are numerical attributes.

2.1.5. Auxiliary Dataset: Secondary Schools

The auxiliary dataset about the secondary schools provides the ***name***, ***latitude***, and ***longitude*** of 150 unique secondary schools. The ***name*** is categorical attribute, and the ***latitude*** and ***longitude*** are numerical attributes.

2.1.6. Auxiliary Dataset: Hawker Centers

The auxiliary dataset about the hawker centers provides the ***name***, ***latitude***, and ***longitude*** of 114 unique secondary schools. The ***name*** is categorical attribute, and the ***latitude*** and ***longitude*** are numerical attributes.

2.1.7. Auxiliary Dataset: Train Stations

The auxiliary dataset about the train stations provides the ***name***, ***codes***, ***latitude***, ***longitude***, ***opening_year***, and ***type*** of 166 records of train stations. However, four of the data points have duplicate names compared to other data points. Upon inspection, we found that the duplicate data points simply had the same name but were recorded separately for different train lines. We will not perform any preprocessing on the duplicated data since we consider the number of lines as a weight when performing integration in Section 2.2. For example, the Paya Lebar station is an interchange station along the East-West Line (EWL) and Circle Line (CCL), it will be considered as two weighted train stations. The ***name***, ***codes***, and ***type*** are categorical attributes, and the ***latitude***, ***longitude*** and ***opening_year*** are numerical attributes. There are 40 records for train stations with missing values for the ***opening_year*** attribute, and one record with an ***opening_year*** in the future. We removed the record with the future ***opening_year*** but retained the other records with missing ***opening_year***. As for the attribute ***type***, there are only two distinct values, which we considered as variants of train based on domain knowledge. Therefore, the attribute ***type*** will not be used when we perform integration in Section 2.2

2.1.8. Auxiliary dataset: Population Demographics

The auxiliary dataset about the population demographics provides the ***planning_area***, ***subzone***, ***age_group***, ***sex,*** and ***count*** of 7836 records. All attributes, except for ***count***, are categorical attributes, while ***count*** is a numerical attribute. Due to the large number of records in this dataset, we merged the records based on ***planning_area*** and ***sex*** attributes.

2.2. Feature Engineering with the Auxiliary Datasets

Based on the analysis and understanding of various auxiliary datasets in the previous Section 2.1, we decided to merge them into the main dataset by creating new features in this section using different feature engineering strategies.

For commercial centers, we considered the number of different types of commercial centers within 3 kilometers and 5 kilometers of the housing location as important features. Due to the limited number of commercial centers in Singapore (38 in total), we chose a relatively large reference distance between the locations of the housing and commercial centers. On the other hand, we considered the types of commercial centers, as different types of commercial centers may have varying impacts on housing prices. As for how to measure distance, we calculated the Euclidean distance using longitude and latitude, and we represented each 1km as approximately 0.009 units based on longitude and latitude.

Primary schools in Singapore have an admission policy that prioritizes placement for registered students, where proximity to the school is a key consideration. Typically, approximately 1km from the school is considered an important threshold in this policy. Therefore, we considered the number of primary schools within 1 kilometer and 3 kilometers of the housing location as important features. Similarly, we measured the number of secondary schools, shopping centers, hawker centers, and train stations with weighted train lines within 1km and 3km radius of the housing locations, as they are relatively evenly distributed across Singapore. The new column names consist of the abbreviation of each additional dataset and a suffix indicating the range. For instance, "the number of primary schools within 1 kilometer" is denoted as ***pri_scl_1***. It is worth noting that for commercial centers, each ***type*** becomes a new attribution. For example, "the number of commercial centers (IHL) within 3 kilometers" is represented as ***com_IHL_3***.

Population demographics is different as compared to other auxiliary datasets. Previously, we preprocessed the population dataset by merging records on ***planning_area*** and gender (***sex***). In this step, we used the corresponding population counts of different genders within each planning area as important features. However, given that the relationship between housing resale prices and population density is stronger than that between housing resale prices and population count, and since we do not have additional information to compute population density, we estimated relative population density by dividing the population count by the number of subzones in the planning area. The intuitive idea is that larger areas tend to be divided into more subzones, and by dividing by the number of subzones, we can simulate population density to some extent. The new column names for male and female are denoted as ***pop_m*** and ***pop_f.***

2.3. Distribution of Variables

After finalizing features in the main dataset, we visualized the distribution of variables to understand the patterns of features.

2.3.1. Target Variable – Resale Price

From the histogram of the target variable in Figure 1, we can observe that resale price is right skewed with a skewness of 0.99.

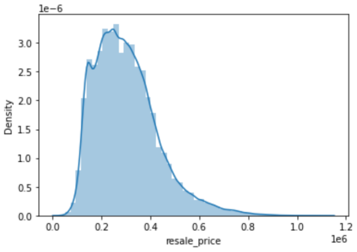

Figure 1: Histogram of Resale Price

2.3.2. Continuous Variables

We plotted the histogram of the continuous variables to visualize the patterns in their distribution. In Figure 2, we can observe that most of the variables do not follow a normal distribution.

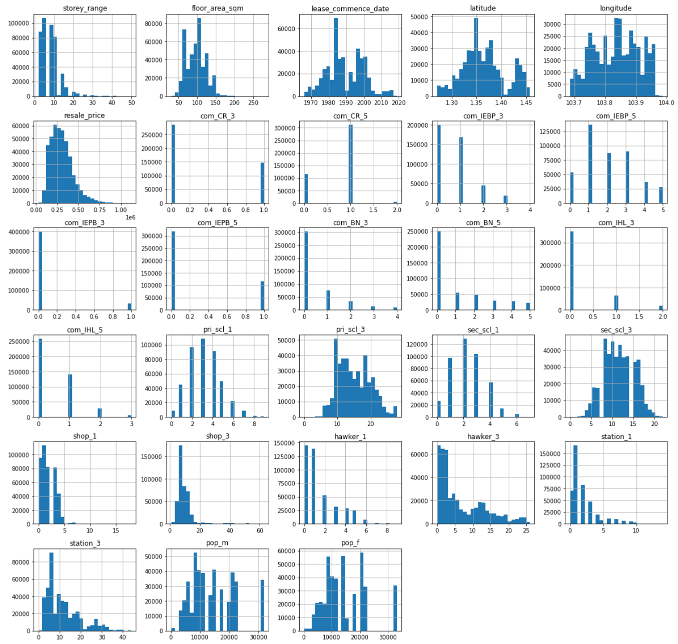

Figure 2: Histogram of Continuous Variables in the Main Dataset

2.3.3. Discrete Variables

We plotted bar charts to visualize the counts of discrete variables by class and some examples are shown in Figure 3. We can observe that some categorical variables are imbalanced with more houses in the west region, most ***flat_type*** of 4 rooms and most ***flat_model*** being “model a” and ‘improved’.

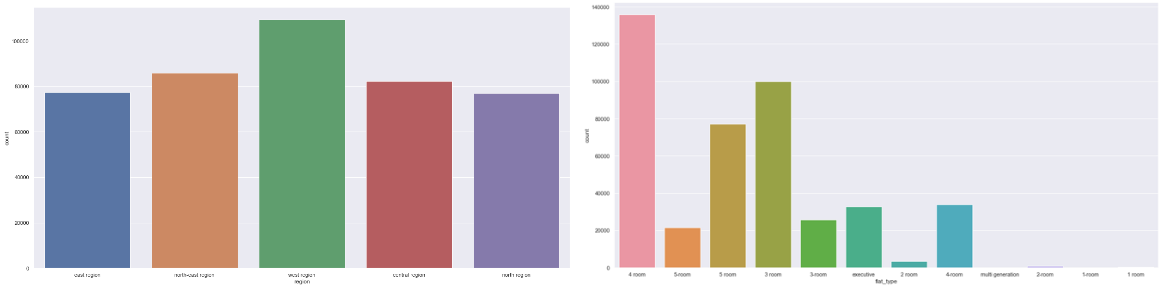

Figure 3: Examples of Bar Charts of Discrete Variables

2.4. <u>Relationship of Resale Price with Variables</u>

2.4.1. Continuous Variables

To analyze the relationship of the target variable - resale price with the continuous features, we plotted the scatter plots of continuous variables on x-axis and resale price on y-axis and some examples are shown in Figure 4. It can be observed that ***storey_range***, ***floor_area_sqm*** and ***lease_commence_date*** have a positive relationship with the resale price.

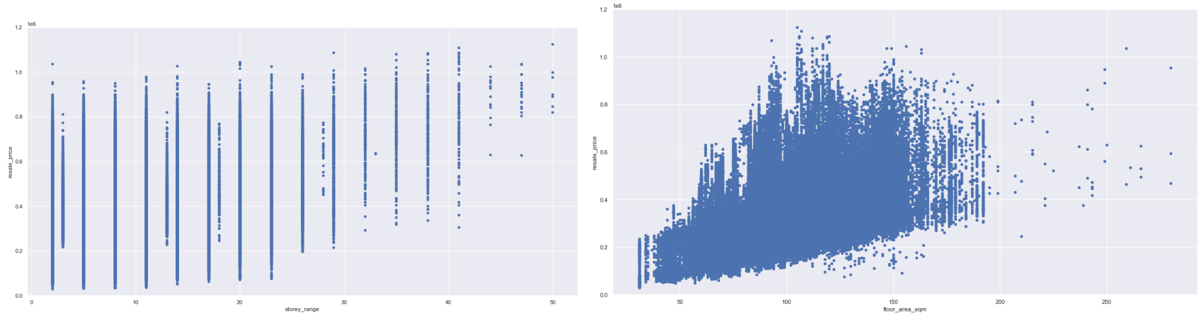

Figure 4: Examples of Scatter Plots of Continues Variables with Resale Price

2.4.2. Discrete Variables

Similarly, to understand the relationship of discrete variables with the resale price, we plotted the boxplots, and some examples are shown in Figure 5. It can be observed that for some discrete variables, certain classes have a higher or lower resale price. For example, for ***flat_type***, ‘multi generation’ have a much higher values for resale price whereas ‘1 room’ have the lowest resale price. For region, resale prices of ‘central region’ vary more than others.

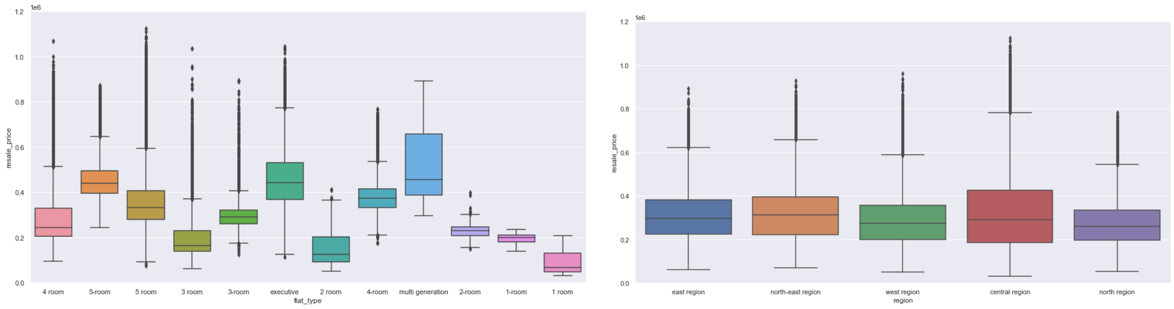

Figure 5: Examples of Box Plots of Discrete Variables with Resale Price

2.5. <u>Data Correlation Analysis</u>

To analyze the correlation of continuous variables among them and with the target variable resale price, we plotted the heat map shown in Figure 6. It can be observed that ***sec_scl_1*** (number of secondary schools within one kilometers) has 0 correlation with ***resale_price***, indicating that there is almost no correlation between this feature and the target variable. Thus, we dropped ***sec_scl_1*** for model development.

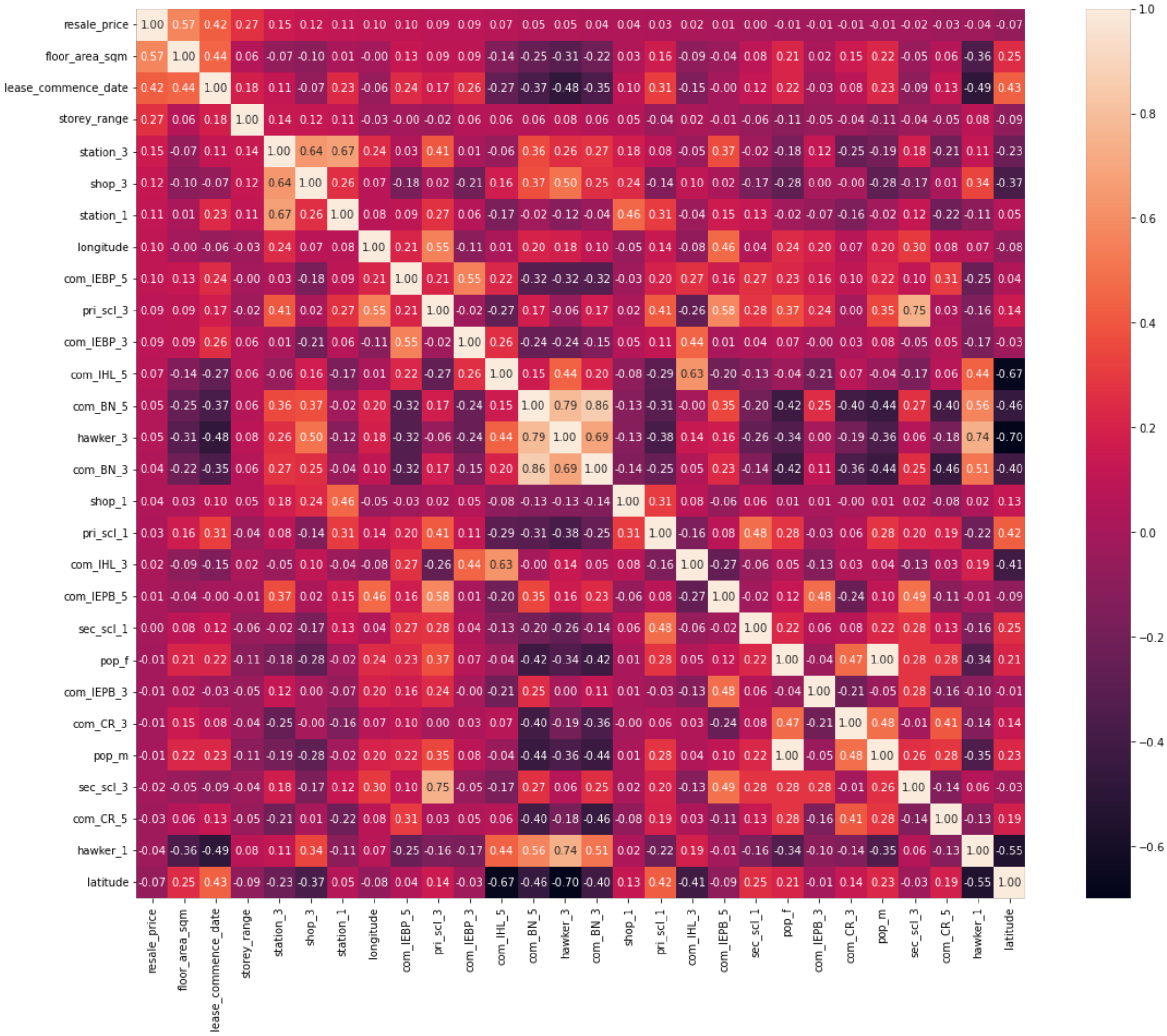

Figure 6: Heat Map of Correlations among Continuous Variables

2.6. Location Map Analysis

As the main dataset contains information about latitude and longitude, we plotted the location map of variables to visualize their relationship in Singapore, and some examples are shown in Figure 7. It can be observed that resale prices are higher at the west and central regions with more darker points clustered around. Number of train stations within three kilometers (***station_3***) have a similar pattern with more darker points in the west, north-east and central regions.

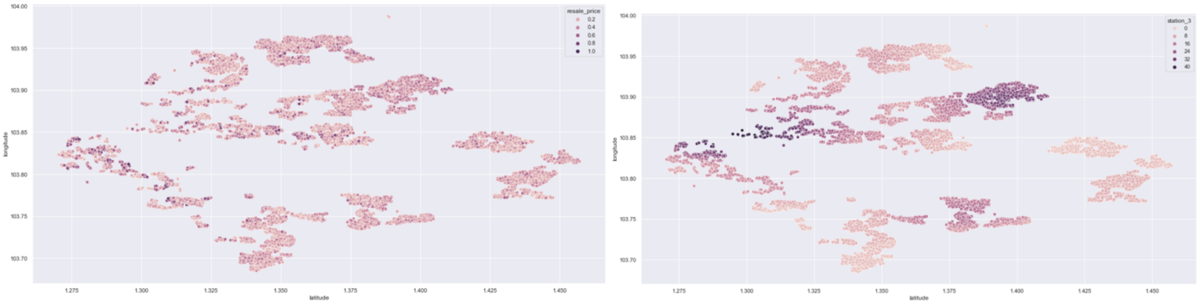

Figure 7: Location Map of Variables, ***resale_price*** (Left), ***station_3*** (Right)

2.7. Dropping Features

Besides dropping some continuous variables based on the correlation in Section 2.5, we also dropped some categorical variables which have a high correlation with each other. For example, as ***region***, ***town*** and ***subzone*** have a high correlation with ***planning_area*** and provide similar information about the location of the house, we kept ***planning_area*** only to represent these columns dropped. We dropped ***flat_type*** and kept ***flat_model*** as these two variables is also highly correlated. Furthermore, we also dropped unnecessary columns which are not useful in resale price prediction such as ***street_name*** and ***block***. ***street_name*** is hard to process as it is in text and the location information can be represented by ***latitude*** and ***longitude***.

2.8. Handling Categorical Data

We changed ***month*** into numeric variable, for example from '2010-08' to 201008. Furthermore, we performed target encoding on the categorical variables ***planning_area*** and ***flat_model*** as they contain 32 and 20 unique values, respectively. One-hot encoding would significantly increase the number of attributes for the current set of approximately 30 attributes. Therefore, we deemed target encoding to be a more appropriate choice than one-hot encoding.

3. Data Mining Methods & Evaluation

To build prediction models for resale prices, our target variable is ***resale_price***. For independent features, we have 30 features including ***month***, ***planning_area***, ***flat_model***, ***storey_range***, ***floor_area_sqm***, ***lease_commence_date***, ***latitude***, ***longitude***, ***com_CR_3***, ***com_CR_5***, ***com_IEBP_3***, ***com_IEBP_5***, ***com_IEPB_3***, ***com_IEPB_5***, ***com_BN_3***, ***com_BN_5***, ***com_IHL_3***, ***com_IHL_5***, ***pri_scl_1***, ***pri_scl_3***, ***sec_scl_1***, ***sec_scl_3***, ***shop_1***, ***shop_3***, ***hawker_1***, ***hawker_3***, ***station_1***, ***station_3***, ***pop_m*** and ***pop_f***.

3.1. Methods & Evaluation

We split the given training dataset (train.csv) into a new training set and a validation set using the train_test_split function with a ratio of 8 to 2. We performed hyperparameter tuning using the GridSearchCV function with the default 5-fold cross-validation to obtain the best combination of hyperparameters for each model. Then, we compared the performance of different models with their optimal set of hyperparameters on the same validation data and evaluated based on the root mean square error (RMSE). The experimental results demonstrate that the **Catboost** model yielded the lowest error, which means the predictions were closest to the ground truth values.

The performances of the different models, along with their corresponding best combination of hyperparameters are show in Table 2 below.

Model	**Hyperparameters**	**RMSE on validation set**
Decision Tree	'max_depth': 25, 'min_samples_split': 20	22053.55750142075
Random Forest	'max_depth': 25, 'min_samples_split': 20	18605.81355164614
Histogram-based Gradient Boosting	'learning_rate': 0.2, 'max_depth': 20	20740.679672418977
XGBBoost	'learning_rate': 0.1, 'max_depth': 15,	16689.253783018816

	'reg_alpha':0.5	
Catboost	'depth': 15, 'learning_rate': 0.05	16100.720404797878

Table 2: Performances of Different Models and their Hyperparameters

3.2. Discussion

The high performance of the **Catboost** model can potentially be attributed to several factors below:

1. Catboost uses regularization techniques and early stopping strategies, which help prevent overfitting and improve the model's generalization ability.
2. Catboost can detect outliers during the training process and exclude them from the model. This can improve the model's performance by preventing it from being influenced by outliers.
3. Catboost implements robust decision boundaries by using a technique called gradient-based one-sided sampling (GOSS) to prioritize the training of examples that are most likely to result in changes to the decision boundary. By focusing on these examples, CatBoost can build more complex decision boundaries that are able to capture the non-linear relationships between the input features and the target variable while still being robust to outliers in the data.
4. Catboost can automatically handle the categorical features. However, to make the comparison between models fairer, we performed some reasonable data pre-processing steps before training models, and all the features that feed into the models are encoded in numerical values.
5. The unbiased boosting technique in CatBoost is primarily designed to address the issue of bias towards the majority class in imbalanced classification problems, while we are working on the regression problem. However, the technique could potentially be adapted for use in regression problems as well.

4. Conclusion

In this project, we examined the dataset to comprehend its characteristics and potential challenges in data cleaning. The given dataset is relatively clean as there are no missing values in the main dataset. We performed feature engineering to integrate the main dataset with other auxiliary datasets and encoding the categorical features. Furthermore, we performed data analysis to drop some unnecessary features and features having high correlation with each other to ensure our models were trained on high-quality data. We then selected five regression models, including Decision Tree, Random Forest, Histogram-based Gradient Boosting, XGBoost, and Catboost. We assessed each model's performance using cross-validation and tuned their hyperparameters to optimize their results. After comparing the RMSE of all models, the experimental results demonstrate that the Catboost outperforms other models, and we provide potential reasons for its outstanding performance. This project with an accurate prediction of resale prices of the houses in Singapore can help Singaporeans to evaluate the resale prices being overpriced or underpriced so as to make a better decision.