Coursera Capstone Project – The Battle of Neighbourhoods

Table of Contents

1. Introduction

1.1. Scenario and Background

I am a data scientist currently working in a Property Agency in Singapore. My boss, the stakeholder would like to know the flat price given the address of property, he wants me to predict the flat price without knowing other characteristics such as flat type, area, remaining lease and etc.

He also mention that the flat price may be highly related to the environment of the house. Generally, the living would be more convenient if your house is near the metro stations, schools or shopping malls, therefore, the flat price near the more venues would be higher. Moreover, the distance to downtown is also significant effect for house price, as the price will be higher for the properties located nearer downtown.

In this project, I will apply the learned skills during the Coursera course to explore the venues around house block address as features of training data and applied some regression algorithms to build the predictive models to solve this problem.

1.2. Challenge to Solve the Problem

The challenge to solve this problem is being able to find the venues around given the flat address as features to create the training dataset. Therefore, in order to obtain the all venues, there are some API that will be adopted in solving this problem:

- OneMap API to obtain the coordinates for given flat addresses.
- FourSquare API to obtain the venues for each flat.

1.3. Interested Audience

I believe this is a relevant project for a Property Agency Company or Government House Agency who would like to predict the house price given the house address, thus, they are able to have some sense what is the house price for the flats to be transacted, subsequently, they can advise the reasonable price to their potential house buyers or sellers, since the approach and methodologies used here are applicable in all other places.

Lastly, this project is a good practical case toward the development of Data Science skills.

2. Data Section

2.1. Description of Data

The following data is required to solve the issues:

- House transaction data from 2015 to 2019 which is downloaded from Singapore government datasets website https://Data.gov.sg.
- Coordinates data for each block from OneMap API.
- Venues data for each block from FourSquare.

2.2. House Transaction Data

The house transaction data contains 79100 transaction records from 2015 to 2019, As the house price has been stabilizing during this period in Singapore, this transaction data is suitable for our study. The following table presents the variables contained in the dataset.

Table 1: Breakdown of Data Information within House Transaction Dataset

S/N	Variable Names	Variable Description
1	month	Month of transaction occur
2	town	Town of transacted flat
3	flat_type	Flat type of each transacted flat
4	block	Block no. of transacted flat
5	street_name	Street name of transacted flat
6	storey_range	Storey range of transacted flat
7	floor_area_sqm	Floor area of transacted flat
8	flat_model	Flat model of transacted flat
9	lease_commence_date	Lease commence date of transacted flat
10	remaining_lease	Remaining lease of transacted flat
11	resale_price	Resale price of transacted flat

To get coordinates of each block address for scraping venues data subsequently, the house transaction data has been processed as follows:

- the street name and block no. are combined as new variable 'block address'.
- In order to eliminate the effect of flat area, the transaction rate (price per square meter) for each flat is computed (transaction price / flat area). Since our study concentrates on block level, the median transaction rates at block level are aggregated from flat transaction rates and applied as target variable for our machine learning models training.

The following dataset has been constructed and extracted after processing of house transaction dataset.

	resale_rate	flat_address
0	4250.000000	174 ANG MO KIO AVE 4
1	4044.117647	541 ANG MO KIO AVE 10
2	4130.434783	163 ANG MO KIO AVE 4
3	4264.705882	446 ANG MO KIO AVE 10
4	4264.705882	557 ANG MO KIO AVE 10

2.3. Coordinates Data for Block Address

Since our aim is to get the venues based on block address coordinates, It is crucial for us to have the coordinates data of each block as input of FourSquare API.

To achieve this purpose, OneMap API is adopted, OneMap is the authoritative national map of Singapore with the most detailed and timely updated information developed by the Singapore Land Authority. There are also many useful day-to-day information and services contributed by government agencies.

The coordinates data obtained from OneMap is then merged with resale rate data and shown as below.

	flat_address	resale_rate	LATITUDE	LONGITUDE
0	543 HOUGANG AVE 8	3728.155340	1.37785956993059	103.892020656191
1	331 JURONG EAST AVE 1	3915.175573	1.35082650419633	103.73042009299
2	222 PENDING RD	3408.756868	1.37521334525043	103.774095682531
3	521 ANG MO KIO AVE 5	5617.393720	1.37357054041248	103.851551329495
4	222 SERANGOON AVE 4	4741.320571	1.35842952473613	103.8712240101

2.4. Venues data for Block Address

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. In this project, I will fetch the venues information from Foursquare by feeding in the block coordinates.

For resident houses, the most common venue categories that is associated with living convenience are ‘Food’, ‘Shop & Service’, ‘Bus Stop’, ‘Metro Station’, ‘Arts & Entertainment’ and ‘School’, the corresponding categories ID in Foursquare then found out and applied to get venues.

The first 5 rows of venues data scraping from Foursquare is presented as followings:

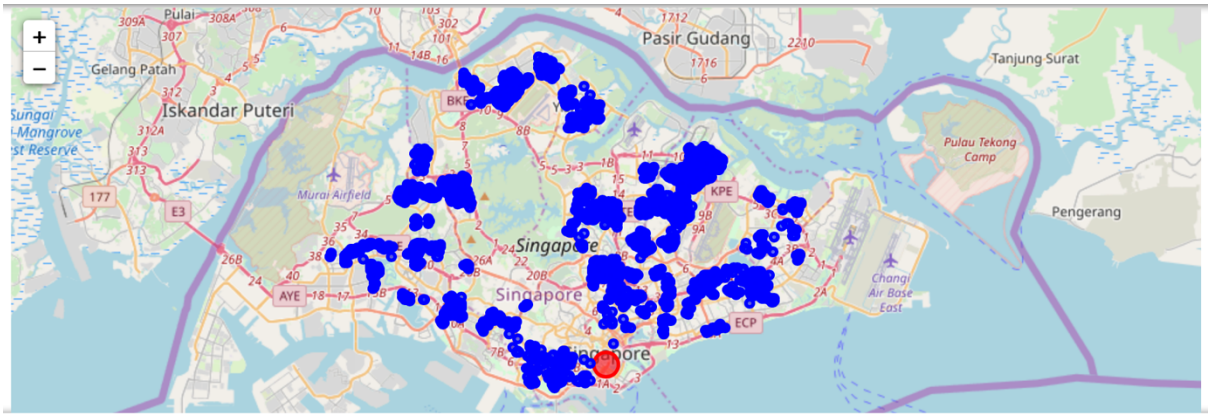
	Flat_Address	Flat_Latitude	Flat_Longitude	Venues	Category	Venue_Latitude	Venue_Longitude
0	543 HOUGANG AVE 8	1.37786	103.892021	21 Food Loft	Food	1.378898	103.887984
1	543 HOUGANG AVE 8	1.37786	103.892021	Fu Fa Food Court	Food	1.379626	103.887903
2	543 HOUGANG AVE 8	1.37786	103.892021	Food Court	Food	1.374471	103.890869
3	543 HOUGANG AVE 8	1.37786	103.892021	A place with food	Food	1.375623	103.891191
4	543 HOUGANG AVE 8	1.37786	103.892021	Tao's Western Food	Food	1.373995	103.891053

Eventually, the venues quantity dataset for each category are built from the venues data, which is capturing characteristics of venues around blocks, below table is the first 5 rows of venue quantity data.

	Flat_Address	Arts & Entertainment	Bus Stop	Food	Metro Station	School	Shop & Service
0	1 CHAI CHEE RD	0.0	4.0	30.0	0.0	5.0	8.0
1	1 GHIM MOH RD	1.0	2.0	14.0	0.0	3.0	7.0
2	1 KG KAYU RD	8.0	3.0	30.0	1.0	6.0	4.0
3	1 QUEEN'S RD	0.0	0.0	8.0	1.0	2.0	2.0
4	10 JLN KUKOH	1.0	3.0	30.0	1.0	6.0	16.0

3. Data Visualization

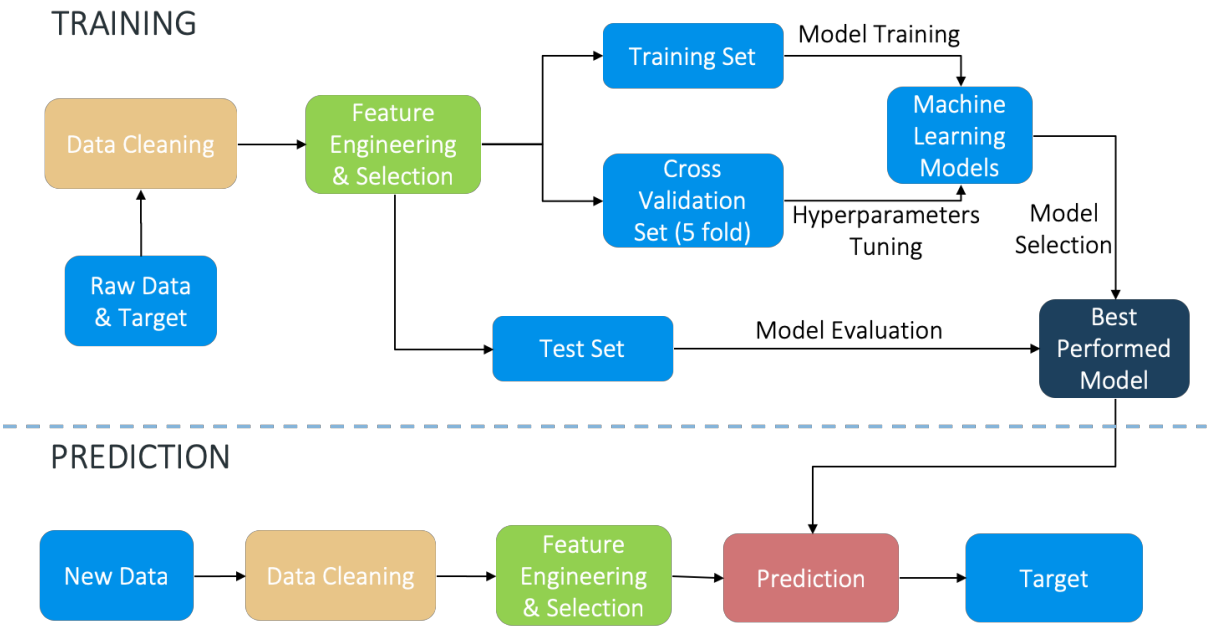
To visualize the block location intuitively, all the block coordinates contained in venue quantity data are passed in 'folium' function and pined to Singapore map. From the graph below, It can be seen that the blocks used in study (blue point) are widely distributed on Singapore main island. They have varied distance to Singapore CBD which is represented as red point in the map below.



4. Methodology

Methodology section represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

For this problem, since the object I study is resale rate given the flat address, I can use regression technique to solve the problem. Below is overview of regression model training work flow.



Prior to building the machine learning models, I have to conduct the feature engineering and apply some data analysis to explore the appropriate features for fitting models in order to gain best model performance.

5. Feature Engineering

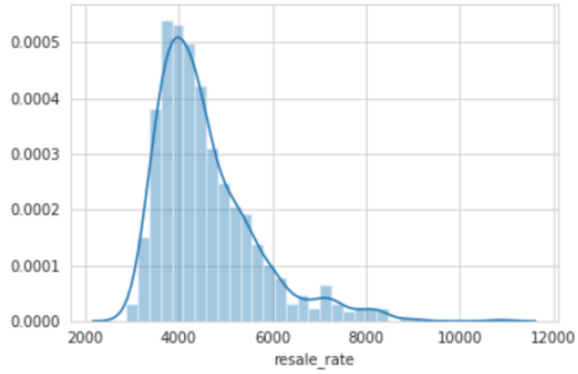
5.1. Feature Creation

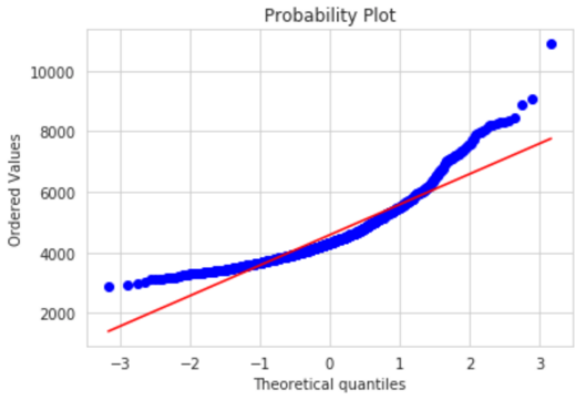
The distance between each block and CDB is deemed as a significant effect for resale price, therefore, it is calculated for each block based on the coordinates as shown below.

	resale_rate	LATITUDE	LONGITUDE	Flat_Address	Arts & Entertainment	Bus Stop	Food	Metro Station	School	Shop & Service	Distance
0	3728.155340	1.377860	103.892021	543 HOUGANG AVE 8	1.0	3.0	9.0	0.0	3.0	8.0	11.405713
1	3915.175573	1.350827	103.730420	331 JURONG EAST AVE 1	0.0	0.0	7.0	0.0	7.0	6.0	15.403301
2	3408.756868	1.375213	103.774096	222 PENDING RD	1.0	2.0	12.0	0.0	1.0	8.0	13.340031
3	5617.393720	1.373571	103.851551	521 ANG MO KIO AVE 5	0.0	3.0	30.0	0.0	9.0	27.0	10.004873
4	4741.320571	1.358430	103.871224	222 SERANGOON AVE 4	0.0	2.0	7.0	0.0	2.0	9.0	8.603372

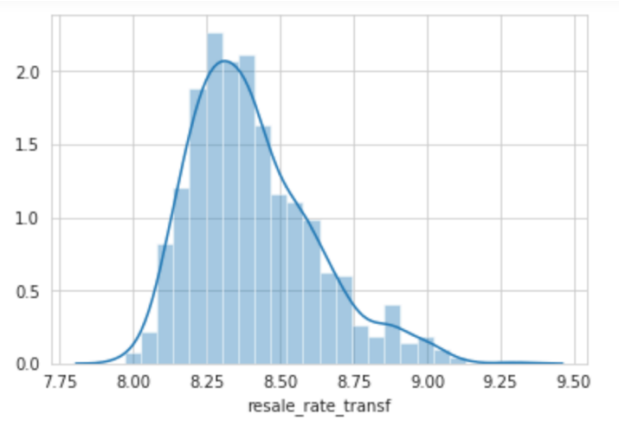
5.2. Data Transformation

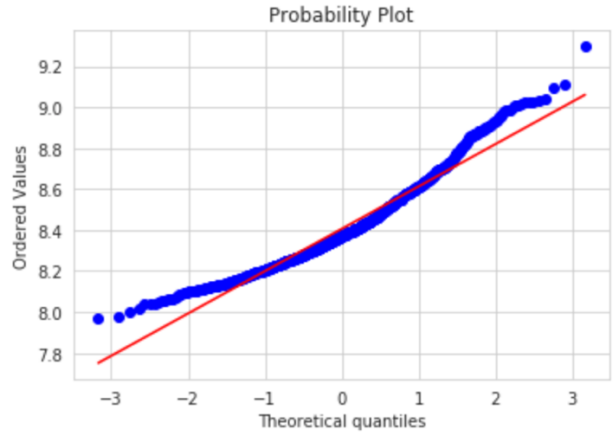
Below graph is the data distribution of the target variable ‘resale_rate’, it is evident that the data distribution is left skewed which might make the trained models biased.





The logarithm transformation is applied to overcome this problem. Let's check the data distribution again, it is more normal distribution now.

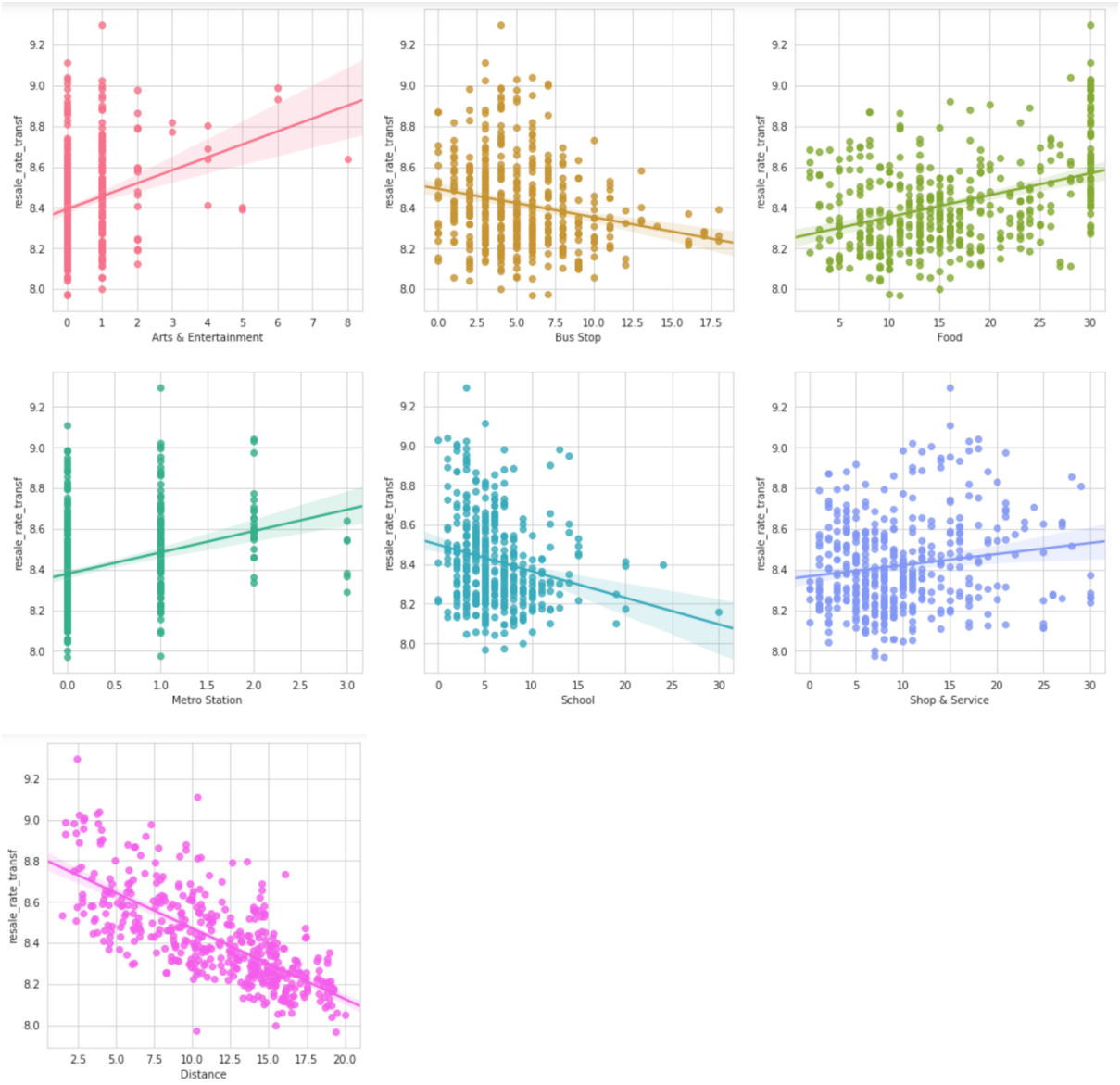




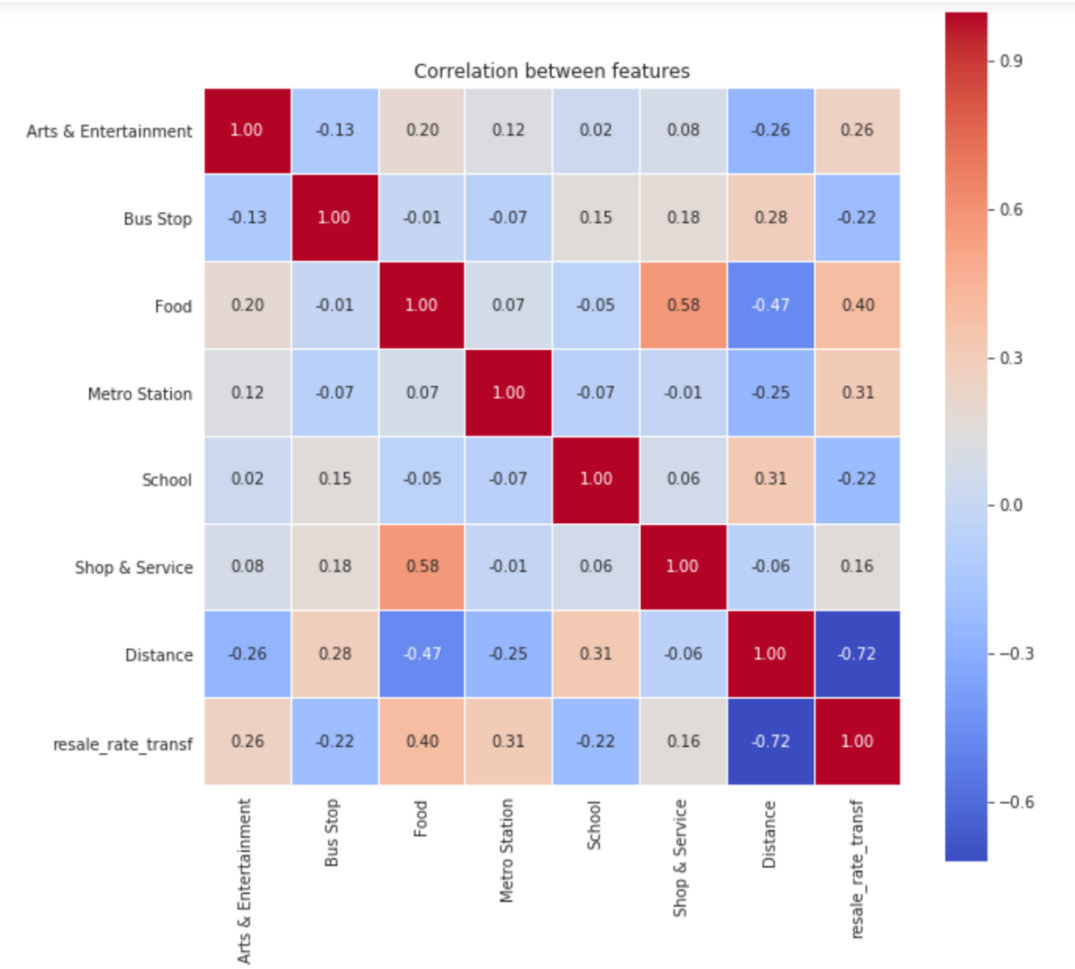
6. Exploratory Data Analysis

Since all features that to be employed in machine learning training are numerical data, the scatter graphs with regression line are plotted to explore the relationships between each features and target variable.

From the figures below, It is found that the correlation between resale_rate and venues are not obvious except distance variable, which is negative correlated to resale_rate.



Correlation between each pair of features and target variable are calculated and plotted as heatmap.




```
Sorted Correlation:
resale_rate_transf      1.000000
Food                    0.398965
Metro Station           0.307887
Arts & Entertainment    0.255902
Shop & Service          0.157386
Bus Stop               -0.215241
School                 -0.219601
Distance               -0.721675
```

Heatmap plot for correlations above show us that any feature pairs (exclusive of target variable) are not highly correlated, which is meaning that there are no redundant features, while every feature is correlating to target variable.

7. Model Analysis

7.1. Feature Selection

As with all machine learnings, the crucial step is to always identify key features which will be used to build our training models. In accordance with the Sec.6.0 above, the features about all venue categories and distance will be selected to fit the machine learning models.

7.2. Train Test Sets Split

The prepared data is split into training set(80% of data) and test set(20%), the training set will be used to train the regression models and do validation and parameter tuning, while test data will be used to model evaluation.

```
#split to training data and test data
X = flat_venues_sub.drop(['resale_rate_transf'], axis = 1)
y = flat_venues_sub['resale_rate_transf']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

print('\nShape of All Sets:')
print(X_train.shape, y_train.shape, X_test.shape, y_test.shape)
```

```
Shape of All Sets:
(720, 7) (720,) (181, 7) (181,)
```

7.3. Model Training and Validation

As the data is very small due to Foursquare API limitations, the following algorithms are selected to train models and evaluated.

- Linear regression
- Decision tree regressor
- Random forest regressor
- XGboost regressor
- LightGBM regressor

Ensemble learning algorithms has proven to outperform a single regressor, XGBoost and LightGBM are widely-used in many winning solutions of machine learning competitions.

For the gradient descent based algorithm such as linear regression, the data shall be normalized prior to fitting models, I will apply StandScaler to normalize the data.

With tuning of the hyperparameters of each models, the final prediction mean square error for the 5 different machine learning techniques are summarised in the table below:

Machine Learning Technique	**Mean Square Error**
Linear regression	0.0202
Decision Tree	0.0244
Random Forest	0.0128
XGBoost	0.0148

LightGBM	0.0145

Among the all techniques, Random Forest returns the lowest mean square error for predicting the resale rate, while Linear Regression and Decision Tree performed worse.

7.4. Model Evaluation

The best performed model of Random Forest with hyperparameters are then evaluated by using the test data.

From model evaluation results shown below. The average error is only 417.67 dollars/m^2, and accuracy is up to 0.915 compared to average resale rate which is 4573 dollars/m^2. This means that our prediction algorithm was successful in predicting the resale rate given a new flat address.

```
'RF' Model Performance
Mean Squared Error(log transformed): 0.0149.
Average Error: 417.67.
Average Resale Rate: 4572.7.
Accuracy = 91.52%.
```

8. Discussion

The evaluation results proof that the model trained is able to perform very well for out of samples (new data).

However, we have to admit that the data utilized in the models training is too small to reflect some insights, and that may make the models trained biased. In order to improve our model performance, more venue data shall be collected from Foursquare.

Furthermore, more time should also be spent on exploration of more venue categories such as entertainment place, wet market(specific market in Singapore), and etc.

Lastly, with expanding of our training dataset, the Artificial Neural Network could be explored and applied in this project to make performance better.

9. Conclusion

The problem of house price prediction is one of crucial problem for the private property agency and government house agency. With the house price predicted, stakeholders would carry out comparison with actual price to check whether the price is reasonable or not, or suggest most reasonable house price to potential flat buyer or seller to fulfil customer's demands, consequently improve business profitability.

In this project, I have conducted an effective feature engineering technique by applying Foursquare API to obtain the venue categories and corresponding data. With combination with another new created feature, i.e. Distance, the training dataset is then constructed to fit the regression model. In the model training, validation and evaluation sections, it can be seen that the model trained achieved very good performance. Therefore the stakeholder's problem would be solved with deployment of this model.