Coursera Capstone Project – The Battle of Neighbourhoods

Outline

- Introduction
- Data Section
- Data Visualization
- Methodology
- Feature Engineering
- Exploratory Data Analysis
- Model Analysis
- Discussion
- Conclusion

1.
Introduction

Project Objective

- The stakeholder would like to know the flat price given the address of property.
- To predict the flat price without knowing other characteristics such as flat type, area, remaining lease and etc.
- Apply the learned skills during the Coursera course to explore the venues around house block address as features of training data.

2.
Data Section

Dataset Description

The following data are required to solve the issues in analysis:

- 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.

House Transaction Dataset

- Contains 79100 transaction records from 2015 to 2019
- 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

House Transaction Dataset

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).
- 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

Coordinates Data for Block Address

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

Venues Data for Block Address

- 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 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

Venues Data for Block Address

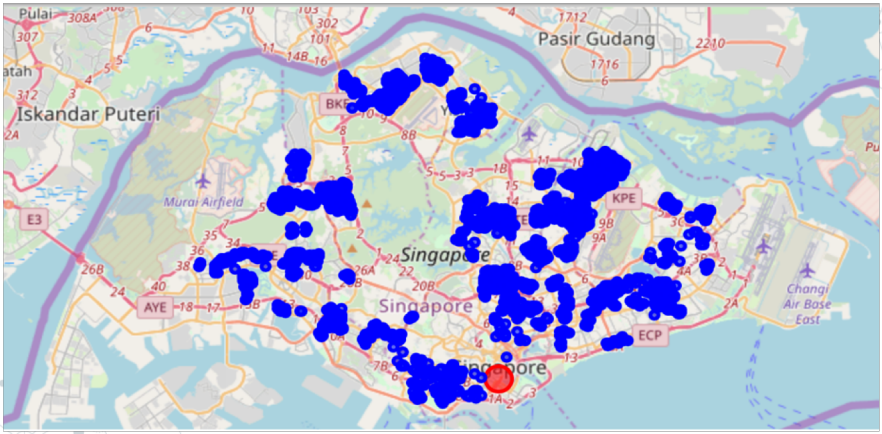
Eventually, the venues quantity dataset for each category are built from the venues data, which is capturing characteristics of venues around blocks.

	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

Data Visualization

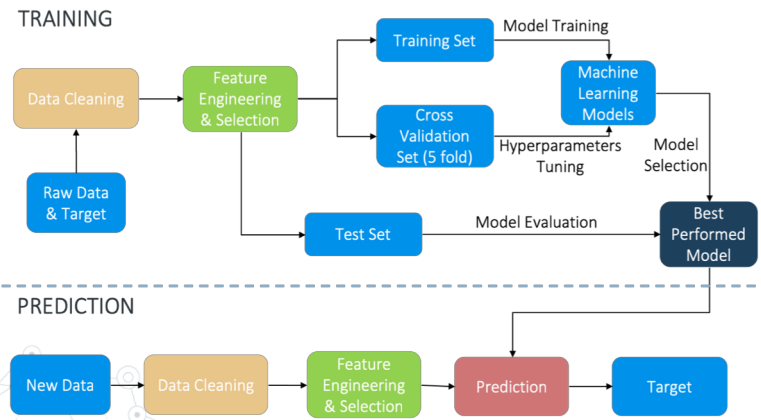
- All the block coordinates contained in venue quantity data are passed in ‘folium’ function and pined to Singapore map.



4.
Methodology

Methodology

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 process.



5.
Feature Engineering

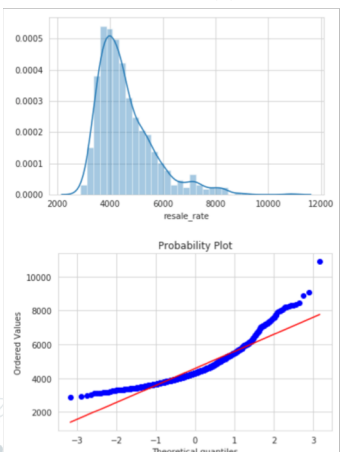
Feature Creation

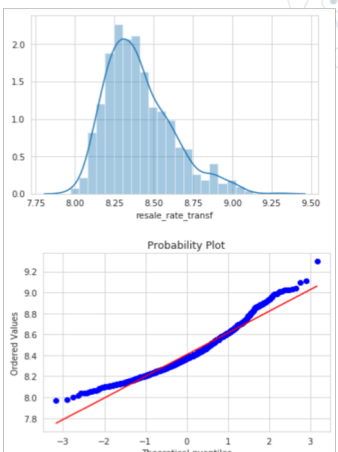
The distance between each block and CDB is deemed as a significant effect for resale price, therefore, it is calculated for each block based on the coordinates as shown below.

	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

Data Transformation

The data distribution of ‘resale_rate’ variable is left skewed which might make the trained models biased. The logarithm transformation is applied to overcome this problem

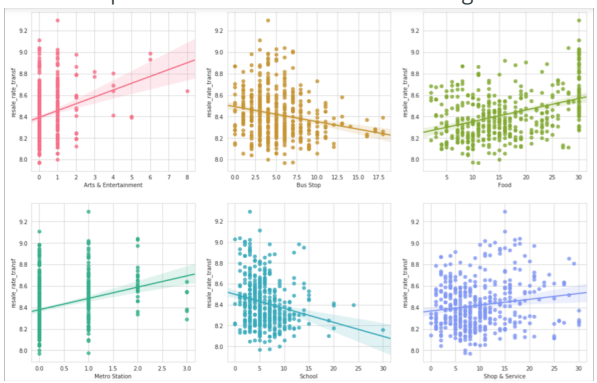




6.
Exploratory Data Analysis

Exploratory Analysis

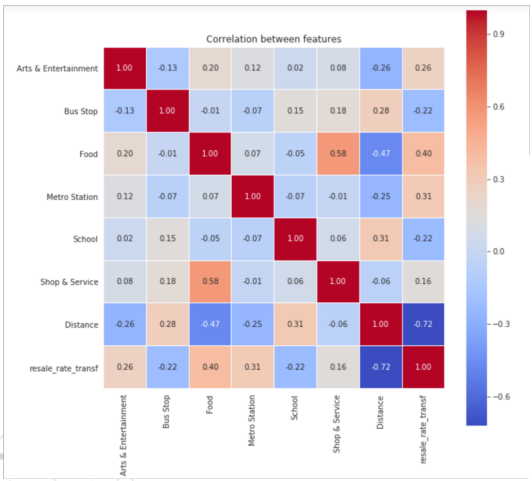
the scatter graphs with regression line are plotted to explore the relationships between each features and target variable.



- Correlation between resale_rate and venues are not obvious.
- Distance variable is negative correlated to resale_rate.

Exploratory Analysis

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
```

- Any feature pairs (exclusive of target variable) are not highly correlated, which is meaning that there are no redundant features.
- Every feature is correlating to target variable.

7.
Model Analysis

Model Training and Validation

- 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.
- The prepared data is split into training set(80% of data) and test set(20%).
- The following algorithms are selected to train models and validated.
  - Linear Regression and Decision Tree
  - Tree based ensemble learning algorithms, Random Forest, XGBoost and LightGBM are applied for fitting models.
  - Ensemble learning algorithms has proven to outperform a single regressor.
  - XGBoost and LightGBM are widely-used in many winning solutions of machine learning competitions.

Model Training and Validation

- Each trained model is 5-fold cross validated.
- Cross Validation Results:

Machine Learning Technique	Mean Square Error
Linear regression	0.0202
Decision Tree	0.0244
Random Forest	0.0128
XGBoost	0.0148
LightGBM	0.0145

Model Evaluation

- Random Forest has been evaluated by test set.
  - Mean Square Error: **0.0149**
  - Average Error: S$ **417.67** /m^2
  - Average Resale Rate: S$ **4572.7** /m^2
  - Accuracy: **91.52%**
- Conclude that our prediction algorithm is able to successfully predict the resale rate for out of samples.

8.
Discussion

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 Neural Network could be explored and applied in this project to make performance better.

8.
Conclusion

Conclusion

- 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.

Thanks!