BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

62nd ISI World Statistics Congress
ISI 2019 Kuala Lumpur

62nd ISI WORLD STATISTICS CONGRESS 2019

18 – 23 August 2019, Kuala Lumpur

Come | Connect | Create

IPS-216

PROXIMITY TO URBAN RAIL TRANSIT (URT) AND ITS IMPACT ON HOUSING PRICES

Goh Jiun Shyan
Bank Negara Malaysia

23rd August 2019, 2.15 pm – 2.45 pm

Outline/Content

1. Objective/Motivation
2. Overview of Data
3. Data Processing and Engineering (Location Features)
4. Model Building
5. Comparison of Model Performance
6. Features Importance
7. Conclusion
8. Suggestion of Improvements

Objective/Motivation

Objective

To determine if proximity of urban rail transit (URT) has great impact on housing prices (Kuala Lumpur and Selangor) using features importance by Random Forest.

Motivation using web-based data (StarProperty)

- Timeliness

Official data lag by at least a quarter

- Granularity

Web-based data have more granular breakdown of address and hedonic features

Overview of Data

StarProperty
(2013 - 2018)

Data size	5 million records
Data categories	Properties for sale Properties for rent
Data type	1. Location: Address, area, city, state 2. Features: Number of bedrooms and bathroom, size of land and build-up area, house type, furnished/ unfurnished 3. Rental tenure 4. Price: Property selling price, Rental price
Data wrangling	Removed records due with missing data, data errors and non-Malaysian addresses. Cross-checked with database of postcodes and utilized Google Maps API to enrich locational information Removed extreme outliers
Final Data Size	500,000 records
Limitations	Listing are heavily skewed to high-rises and upper-middle range properties. Sparse data issue in certain states Limited information on property type

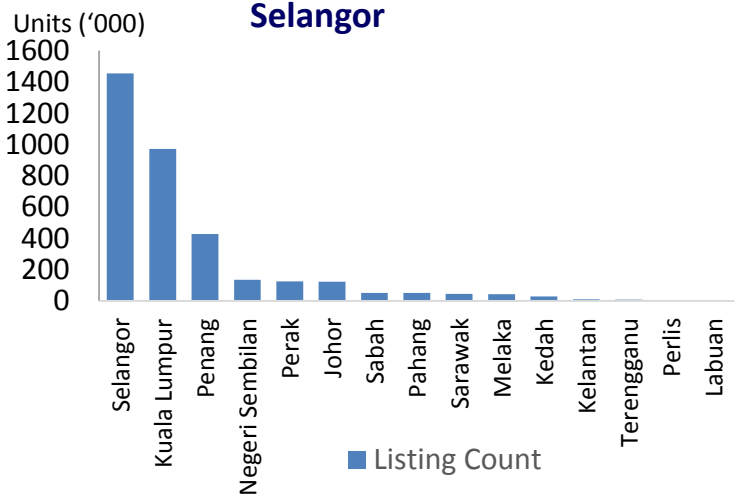
Note: Data preparation (cleaning & organising data) accounts for about 80% of the work of data scientists (source: Gil Press (2016), 'Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says", Forbes).

StarProperty: Profile of domestic market

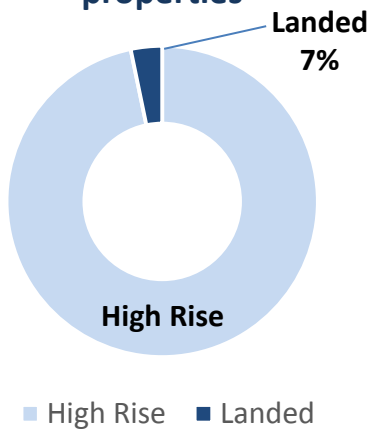
Observations on locations and property types:

1. 68% of properties concentrated in KL & Selangor
2. 93% of listings are for high rise properties

Properties are mostly in KL & Selangor



High rises make up 93% of properties



Data Processing and Engineering (Part I)

Leveraging on External Data to Obtain Additional Information

Objective: Classify **addresses** into **state** and **district**

Addresses		**State**	**District**
Troika, KLCC		Kuala Lumpur	Bukit Bintang
Jalan Kukuban, Taman Setapak		Kuala Lumpur	Wangsa Maju
Jalan Anggerik, Taman Seksyen 10		Selangor	Petaling Jaya

Regular expression with external data (list of states and district)

But, some addresses given lacks information on state and district

Addresses ???
Prima PV20 Apartments
D'Casa Residence

Solution: GoogleMaps

Data Processing and Engineering (Part II)

Leveraging on GoogleMaps API to Obtain Additional Information

Main Idea: Get **longitude and latitude** from addresses

Get longitude and latitude

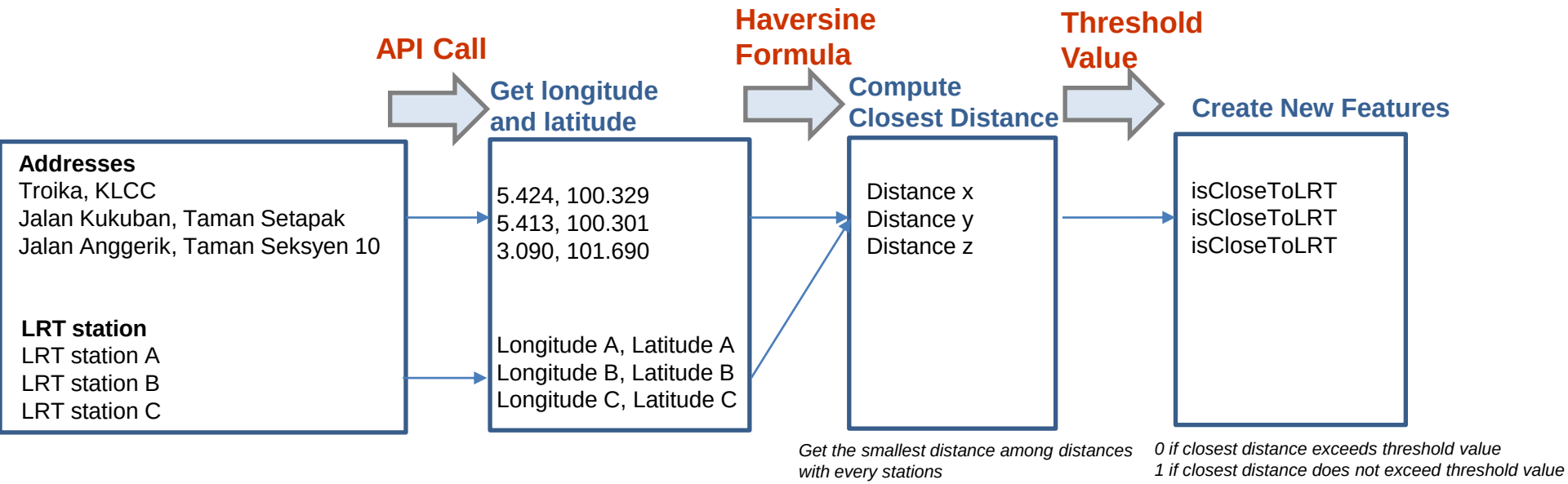
Addresses
Troika, KLCC
Jalan Kukuban, Taman Setapak
Jalan Anggerik, Taman Seksyen 10

API Call

Latitude, Longitude pairs (Geolocation)
5.424, 100.329
5.413, 100.301
3.090, 101.690

Data Processing and Engineering (Part III)

Create new features (proximity to URT)



Repeat process for other types of URT like Monorail, KTM and MRT

Model Building

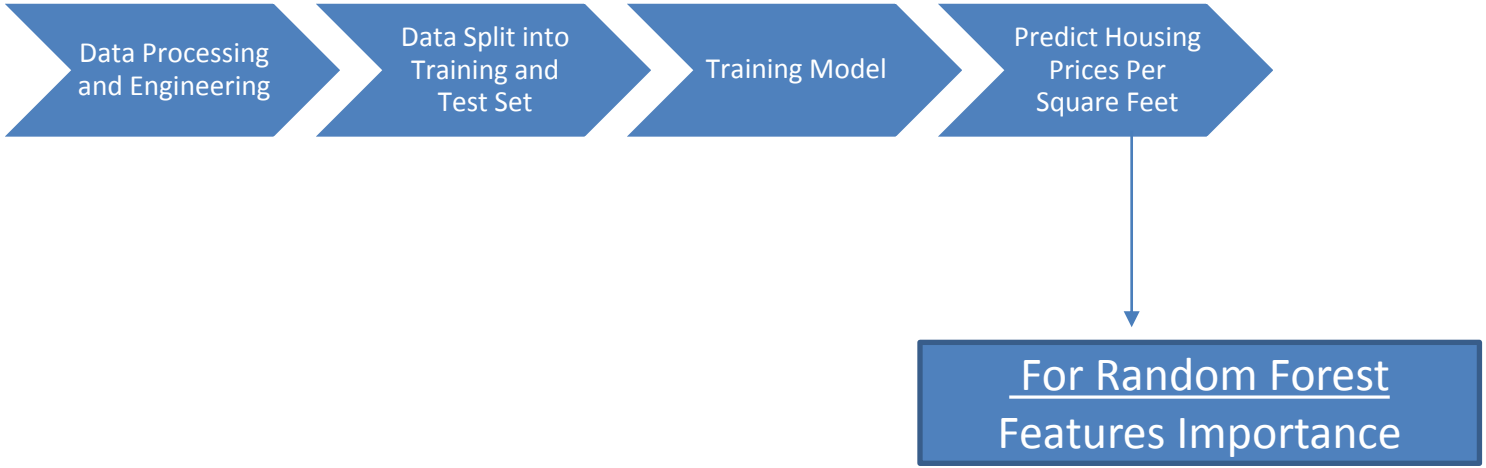
Hedonic Features

- Number of bedroom
- House type
- Furnishing/Unfurnished
- Number of carpark
- Number of bathroom

External Features

- isCloseToLRT
- isCloseToMonorail
- isCloseToKTM
- isCloseToMRT

Model Building



Comparison of Model Performance

Metric/Model	Lasso Regression	Ridge Regression	Random Forest
Hedonic Features Training Error	0.2420	0.2414	0.2324
All Features Training Error	0.2418	0.2412	0.2243
Hedonic Features Testing Error	0.2431	0.2427	0.2363
All Testing Error	0.2430	0.2425	0.2292

Features Importance

Rental Prices

Feature	Gini Score
Number of bedroom	0.2765
Furnishing/Unfurnished	0.1903
House Type	0.1355
Number of bathroom	0.0557
Number of carpark	0.0399
isCloseToMRT	0.0261
isCloseToKTM	0.0244
isCloseToLRT	0.0224

Sales Prices

Feature	Gini Score
Number of bedroom	0.7782
Number of bathroom	0.0700
House Type	0.0191
isCloseToLRT	0.0130
isCloseToKTM	0.0127
Number of carpark	0.0119
Furnishing/Unfurnished	0.0116
isCloseToMRT	0.0090

Conclusion

Considering any new features (isCloseToLRT, isCloseToMonorail and etc) is not top 3 factor ranked by features importance in both sales and rental prices, it is conclusive that proximity of URT does not influence housing prices as much as hedonic features.

Suggestion of improvements

1. StarProperty lists mostly mid-priced houses, other websites like iProperty and PropertyGuru should be used.
2. Threshold value is selected arbitrarily (800m is considered as walking distance). A more scientific approach is recommended.
3. Instead of Gini Impurity, information entropy can be considered to measure feature importance in Random Forest.
4. Other statistical tests/algorithms like ANOVA, genetic algorithm and so can be used to test if URT is an important factor.


62nd
ISI World
Statistics
Congress
ISI 2019
Kuala Lumpur
18 - 23 August 2019 | www.isi2019.org

THANK YOU