CFLD – NUS INNOVATION CHALLENGE

Land Price Evaluation in KUALA LUMPUR

NATIONAL UNIVERSITY OF SINGAPORE
The Elementals

Team Member	**:**	Cho Zin Tun	(A0098996W)
		Peh Yingqi	(A0071186E)
		Tan Yan Zhou	(A0098740W)
		Vinod Vijayakumaran	(A0097740X)
		Wang Jia (Claire)	(A0176605B)
		Wang Shenghao	(A0105772H)
Date of Submission	**:**	February 28, 2018	

Abstract

The use of machine learning to predict land and property prices to make recommendations on investment opportunities presents a more data-driven approach to what is traditionally based on gut-feel and experience. Data on actual transactions that has occurred in the last 10 years was scrapped from online sources and merged with property and land use characteristics, address-level amenities data, and town-level census data to form the dataset. Regression models like LASSO and random forest was trained using the data and the output from the model is further analysed for profits and returns on investment (ROI). Using the profit, ROI, industrial niche and locality information, recommendations the best investment location and type of investment is made.

1. Introduction

This exercise serves as a guideline towards uncovering deeper insights for evaluating raw land prices and realizing viable land areas for potential urban development in a specific city. The objective and the methodology adopted for study is aligned with likes of CFLD's business perspective.

China Fortune Land Development (CFLD), the largest industrial city developer in China is looking into expanding their global presence and Southeast Asian (SEA) countries are of focus, hence this research would provide right direction for strategic decision-making on land viability for development.
To tackle the challenge set forth, multitude of innovative machine learning methods are incorporated for more robust modeling results. These models and results can be further used for future research and development in the space of real estate analytics. Kuala Lumpur (KL) is chosen as the city of focus in this paper among the four other options provided.

Kuala Lumpur is identified to be the second most viable place after Singapore for businesses to operate in, due to its geostrategic location allowing businesses to oversee regional businesses and grow in SEA market. Different municipalities are identified within 50km periphery of KL city, coined as Greater KL for feasibility studies primarily due to the high land cost at the city center. Pro-business government policies like Economic Transformation Programme (ETP) and Invest KL are also introduced promoting businesses to set up their operations around Greater KL region. In this research, 15 regions are identified as focus of study. The research undertaken in this activity backed by model insights would highlight significant factors influencing land prices in different regions within the chosen area. From these results, recommendations are given for optimal investment strategy based on growth potential of regions. Finally, a predictive model is developed to assess future changes in land value for the identified areas.

Since this project places emphasis on data driven decision-making, conscious effort is placed in data sourcing methodology ensuring veracity and variety before inclusion. Careful selection of determinants is done from the available data for the modeling, which makes sure that multiple insights can be drawn. To realize whether the insights provided are substantiated with accurate results, multiple models are valuated with different parameters and results.

Usages of different tools for modeling are highlighted in this paper explaining the platforms through which one could further develop the work done. The foundation of work that has gone into this project would make it useful for research and development purposes.

2. Problem Statement

The main problem to be addressed in this business challenge is to develop methods in evaluating land prices in emerging markets where CFLD is intending to venture. Currently, there is huge untapped market potential from investing in particular regions due to the limited work done in this field to uncover unappreciated areas. This paper aims to address those concerns by establishing a model that explains factors contributing to land value increase and most favourable options for investment.

It is difficult to identify certain countries for future expansion and growth since limited work is done with regards to understanding land and property price fluctuation parameters and to justify a profitable investment strategy. Data that predicts the future land prices are not readily available; hence research is done to tackle this challenge.

This paper explains the approach and methodology adopted in taking the first steps in realizing some of the previously unknown insights. This is done by focusing and researching on a particular city of choice. Certain parameter weighs and contributes differently in each city; hence careful examination is done through cross validation of results from data analytics applications to ensure recommendations given are credible.

3. State of the Art

Property, as a reliable and multidimensional commodity, has a value that is mostly determined by a combination of various characteristics. They are categorized as structural, locational and neighbourhood attributes. Various studies have examined the possible relationships between property prices and distinct attributes through exploring the implicit prices of attributes based on two key methods, the geographically weighted regression (GWR) method and hedonic pricing (HPM) method. Performing analysis work with either methods can easily yield the market prices of specific elements by identifying the corresponding coefficients. Traditionally, properties located within acceptable proximity to commercial centres, green spaces and other facilities commands a greater margin price, as indicated by location theories. Previous studies have also measured 'point of interest' (POI) effects by calculating the key metrics such as distance and/or travel time between the POIs and residentials. However, the existence of such facilities has a limited range effect which varies across different space. In addition to distance considerations, the popularity and relative activity of the space has to be taken into consideration as well. As such, spaces can be classified as either 'cold' or 'hot' spots, depending on the number of visitors as well as the sites' clustering patterns. While such 'hot' spots represent a preponderance of social and/or economic activity, negative impacts such as traffic, noise pollution and security challenges may follow suit too. Currently, few studies have examined the possibility of correlation between the presence of 'hot' spots and property price.

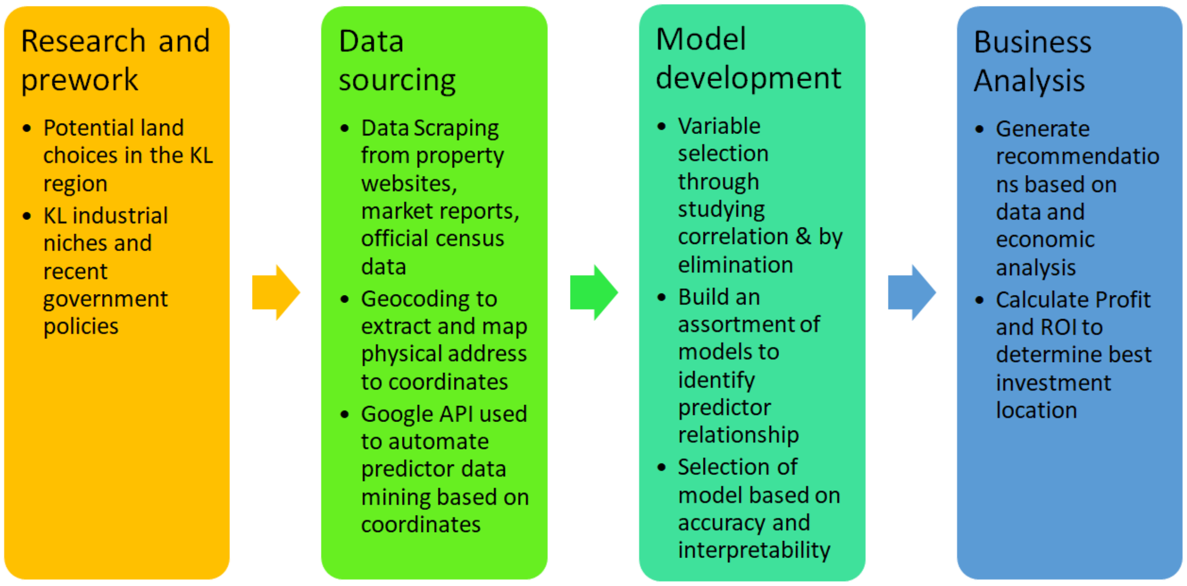
Undoubtedly, implementing the process of collecting statistics on the number of people who access such a POI can be challenging. To overcome such an obstacle, one particular research [2] employed a method to analyse social media data in an attempt to determine the spatial patterns of POI usage and thereafter, the effects on property price. With the widespread usage of social media today, these data can be used as a proxy to big geospatial data. Many business and academic communities have also tapped on the inherent opportunities to study and analyse the urban areas, human behaviour and popular sentiments concerning such areas.

Data analytics in various industries is booming and transforming countless industries. Currently in the property industry, Big Data has empowered various players in optimizing their business strategies based on the analysis of consumer patterns and possible market entrances. As such, the deployment of data analytics can be considered relatively new at this junction. Nevertheless, this paper attempts to utilize the power of current technological tools to aid us in evaluating land prices in emerging markets where CFLD is intending to venture, in particular, Kuala Lumpur. Crawling of a diverse data from credible sources are done automatically through the use of APIs, while the selection of the best plausible model to evaluate land and/or property price in Kuala Lumpur is done with current popular and effective techniques.

4. Solution

4.1 Solution Framework

Our innovative solution consists of the following key segments:



Research and prework

- Potential land choices in the KL region
- KL industrial niches and recent government policies

Data sourcing

- Data Scraping from property websites, market reports, official census data
- Geocoding to extract and map physical address to coordinates
- Google API used to automate predictor data mining based on coordinates

Model development

- Variable selection through studying correlation & by elimination
- Build an assortment of models to identify predictor relationship
- Selection of model based on accuracy and interpretability

Business Analysis

- Generate recommendations based on data and economic analysis
- Calculate Profit and ROI to determine best investment location

4.2 Research on potential land choices and KL region

Based on a 50km radius from the Kuala Lumpur city center, we have shortlisted a total of 15 potential cities/towns. We have excluded Kuala Lumpur city as being the capital of Malaysia, iit is already very well-developed and the possibility to find potential land for future development is not advisable in terms of business value sense. Below is the list of potential cities/towns.

- Ampang Jaya
- Balakong
- Batu Arang
- Beranang
- Bukit Tinggi
- Kajang
- Klang
- Pengkalan
- Petaling Jaya
- Rawang
- Selayang Baru
- Semenyih
- Serendah
- Shah Alam
- Subang Jaya

4.3 Data Sourcing

Based on the domain knowledge and literature review, the following variables are identified as the key predictors of land and property prices.

Category	Predictor	Type	Usability
Time	Year	Numerical	Yes
Census	Population	Numerical	Yes
	Population Density	Numerical	Yes
Economic Indices	GDP	Numerical	No
	CPI	Numerical	Yes
	Inflation rate	Numerical	No
Town data	Township	Categorical	Yes
	Town area	Numerical	Yes
Land property data	Unit cost of construction	Numerical	No
	Property tax	Numerical	No
	Building Type	Categorical	Yes
	Land Area	Numerical	Yes
Geographic data	Geographic Coordinates	Numerical	Yes
	Distance and duration to city center	Numerical	Yes
	No. of miscellaneous amenities	Numerical	Yes
	Distance to nearest amenity	Numerical	Yes

Table 4.1 Predictors of land price

Census

In general, larger population density indicates higher demand of land properties, which further results in higher land and property prices. Population data at town level is available and is used in developing models.

Economic Indices

Economic indices including GDP and CPI reflect the purchasing power of the public. Therefore, GDP and CPI are positively correlated with land and property prices. Since GDP data is not available at town level, only CPI is involved in the land price forecast models.

Geographic data

As is known to all, land property prices vary with geographic locations. In order to examine the effect of geographic locations of land properties quantitatively, the address data is converted to geographic coordinates to represent the locations of land properties. Furthermore, the land price tends to be higher at locations with better availability of infrastructure. The availability of infrastructure can be measured with the no. of amenities such as schools and hospitals within certain distances, as well as the distance to nearest amenity.

4.3.1 Traditional data sourcing on general predictors (population, CPI, etc)

The data of the following predictors are collected from miscellaneous online sources. The state-level CPI data from 2010 - 2017 is extracted from Malaysia's Open Data Portal. The population as well as area data of target satellite cities is collected from the Population List website.

4.3.2 Innovative data scrapping: Python Scrapper, Google Maps API

In this project, the collection of land property data and address-level amenities data is automated with data crawling applications. The land and property data is sourced from a Malaysian property pricing website. The website provides detailed historical transaction data of land and properties at town level. In order to automate the process of data collection, a Python data crawler is developed using Scrapy, a web crawling framework.

In general, the availability of infrastructure around a specific location reflects the level of development in the region. Therefore, satellite cities with better infrastructure development tend to have higher land prices. In order to collect the address-level amenities data, a javascript application is developed with Google Maps API. The address data is converted to geographic coordinates, which is further used to extract the amenities data.

The workflow of collecting amenties data is as follows.

1. Read address data into javascript application;
2. Call Google Maps Geocoding API to convert addresses to latitude/longitude data;
3. Import coordinates data into javascript application;
4. Call Google Places API to extract the geo-information of nearby amenities within certain distance;
5. Calculate no. of nearby amenities within certain distance, and the distance to the nearest amenity and export the results.

4.4 Model Development

Since the business solution was based on the profit, two separate models were implemented. One model was for predicting the empty land price in 2017 while another was for property price in 2020. For the latter model, extrapolated and projected variables were adopted to estimate the predicted property price.

Feature engineering was applied for the missing data and multicollinearity detection was performed to eliminate variables holding the same information.

Various analysis methods, including LASSO regression and random forest, were utilized to build the models. Based on the accuracy and interpretability, random forest was chosen to implement the final model. Moreover, it could provide the variable importance plot, from where key predictors for the property price can be obtained.

4.5 Economic Analysis

The satellite city for real estate investment is proposed based on the expected profits and ROI after the construction is completed and the industrial property is put into use. Since the future data of the key predictors is not available, the following assumptions are made to facilitate the simulation of property price variation.

1. The investment on the real estate project is made in 2017;
2. It takes 5 years to complete the construction. In other words, the industrial estate can be put into use or sold by 2022;
3. Based on the historical average population growth rate, it is assumed that the population of the target towns in Selangor grows by 2.7% annually, while the population of the rest cities grows by 2%;
4. Satellite cities are selected randomly to have their “distances to the nearest school/convenience store” values scaled down by 0.75 randomly.
5. The potential profits and ROI can be estimated with the following procedures.
6. Identify key predictors of property prices: population density, CPI, min distance to school, min distance to convenience store;
7. Estimate projected population of 2018-2022 with historical average population growth rate per year: Selangor 2.7%, the rest cities 2%;
8. Extrapolate CPI of 2018-2022 based on 2012-2017 CPI;
9. Introduce perturbation to min distance to school/convenience store: randomly select the cities with minimum distance (to school/convenience store) larger than certain value (500m / 300m), scale down the min current distance value by 0.75;
10. Calculate the future population density based on the projected population;
11. Predict property prices based on the fitted randomForest model.

12. For a particular year from 2018-2022, the predictors data is derived based on the previous year's predictors;
13. Calculate average fitted property price for each different city;
14. Assume the land is purchased in 2017. The construction of new properties can be completed by 2020. The construction cost is 50% of the land price. Hence, the annual ROI can be computed with the following formula.

$$\text{Annual ROI} = \frac{property\,price - 1.5\ land\ price}{3\ years * 1.5\ land\ price}$$

5. Discussion

5.1 Model Comparison and Analysis

5.1.1 Model selection and comparison

The selection of model was done based on the 2 key criteria of interpretability and accuracy. An assortment of model was tested before the final model is selected. In particular, the LASSO and random forest model will be discussed because of the good fit for the purpose and benefits they confer in this particular use case.

A LASSO model is a regularised form of linear regression. The resulting model is highly interpretable as the variables that are important will have a coefficient and how the variables affect property prices can be deduced from the magnitude and sign of the coefficient. This model also enables variable selection as the regularisation parameter will shrink the variables which are not significant to zero, yielding a parsimonious model that only contains the important parameters.

A LASSO model optimised based on cross validation error for the optimal regularization parameter, lambda, was built using the 27 predictors from the data set. Refer to Appendix A for the LASSO model details.

A random forest is an ensemble model that aggregates different decision trees to produce a highly accurate model. As each decision tree itself is a weak learner, creating an ensemble of trees reduces the overall bias and variance of the model to create a strong learner. Although the random forest is not completely interpretable, the variable importance plot gives a good indicator of how important each variables are in affecting the property prices. Partial dependence plots can also be used to find out the relationship between variable and property prices when required.

A random forest model using 27 predictors was built to model land and property prices. Refer to Appendix B for model details. By optimising the number of variables randomly sampled as candidates at each split of the decision trees within the forest, a cross validation accuracy of 63.92% can be achieved.

Due to the good accuracy and ability to provide information on variable importance and relationship between variable and property prices, random forest was selected as the final model of choice.

5.1.2 Random forest model results

Using the random forest model that was built, variable importance plot was created to visualise the contribution of each variable relative to the other variables, as given in figure 5.1.1 below.

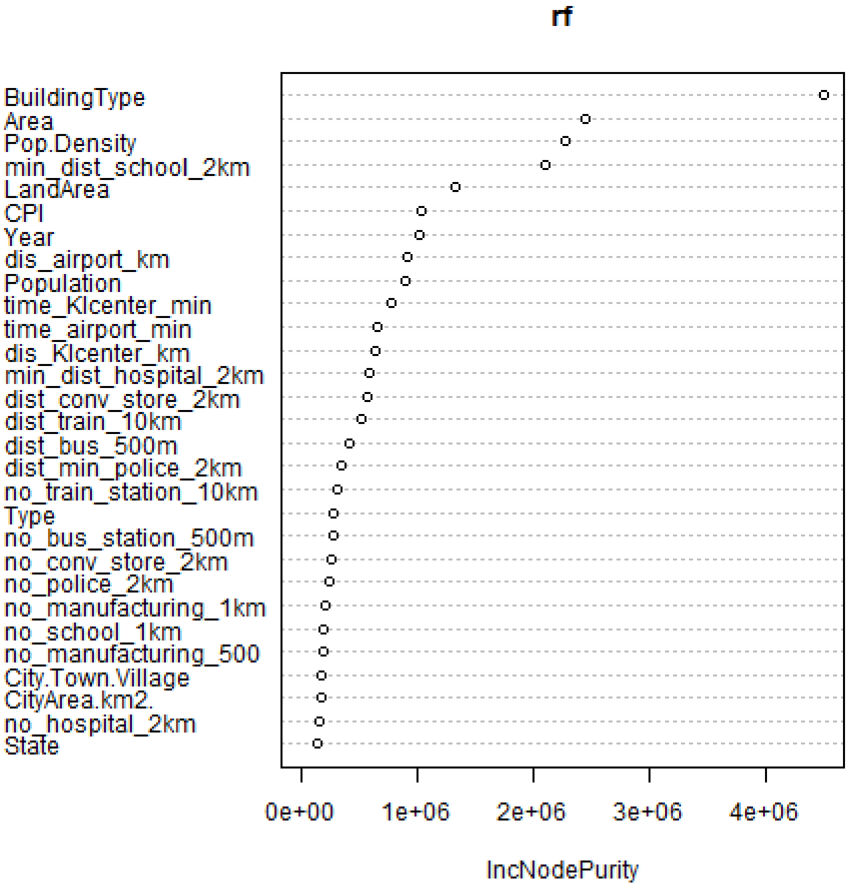


Figure 5.1.1 Variable importance plot of the random forest model

Based on the variable importance plot, we can conclude that the top predictors for property prices are: (1) Land use, (2) Area, (3) Proximity to Amenities such as schools, stores, etc, and (4) Connectedness such as distance to city center, airport, and number of bus stops and train stations in the vicinity.

As the distance to KL city center is an important predictor of land price, we investigate further how land price changes as the distance to KL city center increases. In Figure 5.1.2, the fitted land price per square feet was plotted against the distance to KL city center.

The fitted logarithmic trend line shows the roughly negative relationship between fitted land price in 2017 and distance to KL center. In general, the farther distance from the surrounding city center to KL center, the lower land price in that city. Industrial cities in Selangor are **Subang Jaya, Shah Alam, Klang, Kajang, Rawang, Selayang, Ampang Jaya and Petaling Jaya**. **Port Klang** plays a key factor in the industrial development of Selangor as it is the busiest port in Malaysia. This explains why Kajang, Subang Jaya and Klang are outliers from the trend line.

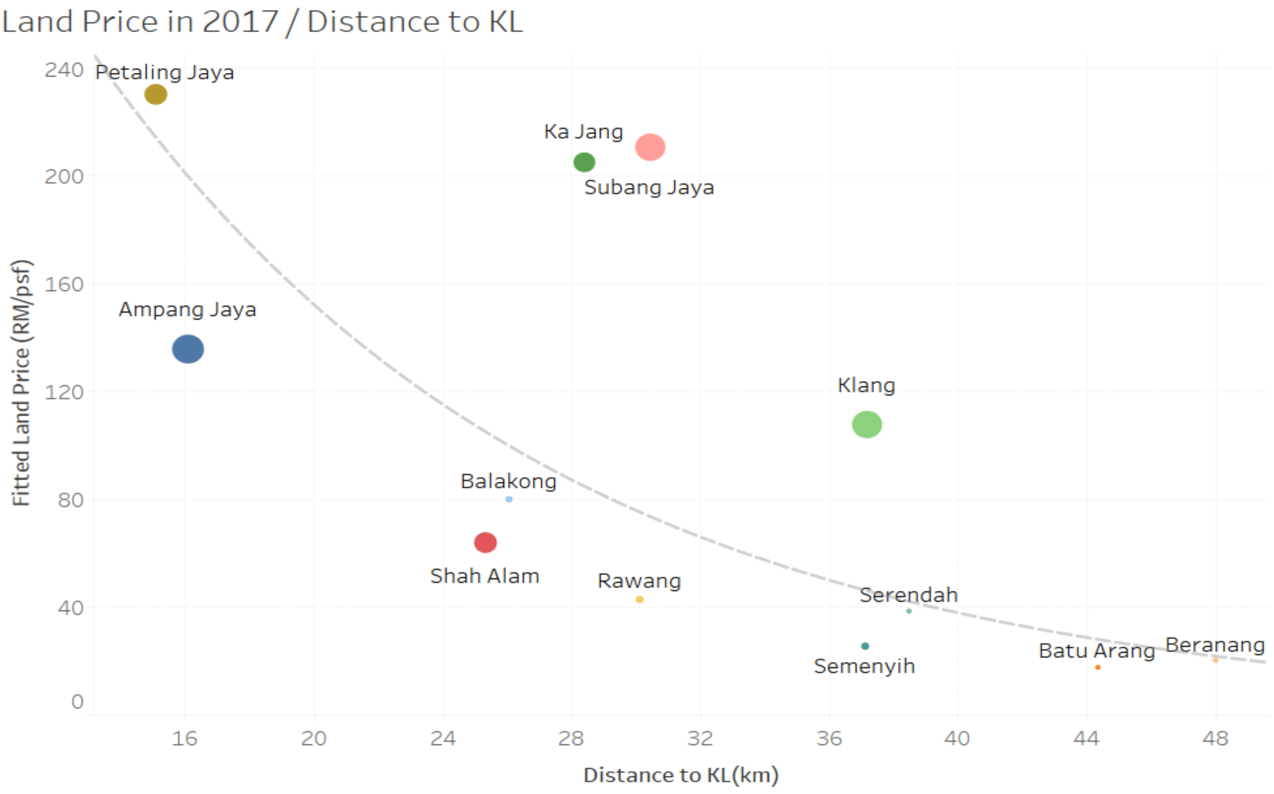
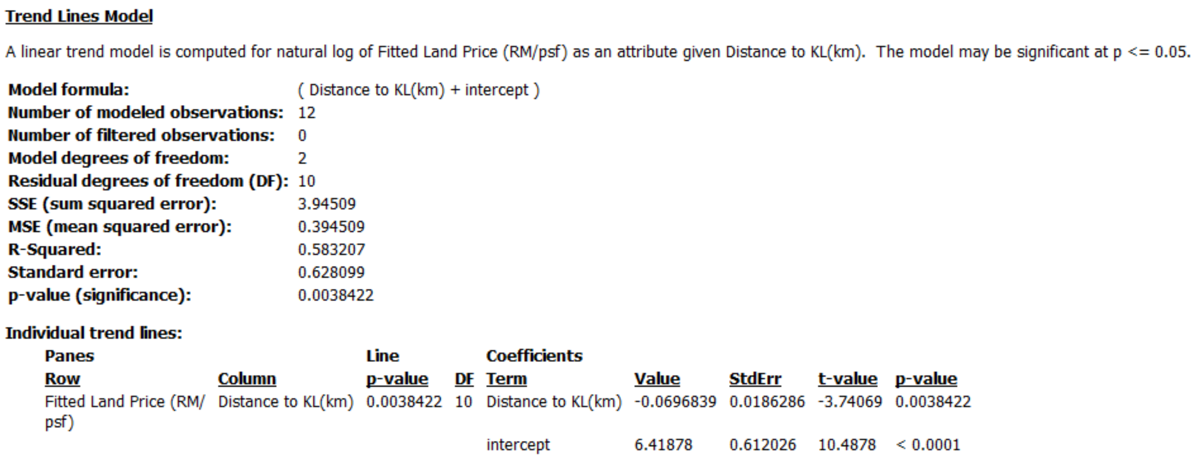


Figure 5.1.2. Land Price in 2017 against Distance to KL by city[1]

[1] The size of bubble denotes population of the city.



Trend Lines Model

A linear trend model is computed for natural log of Fitted Land Price (RM/psf) as an attribute given Distance to KL(km). The model may be significant at p <= 0.05.

Model formula:	(Distance to KL(km) + intercept)
Number of modeled observations:	12
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	10
SSE (sum squared error):	3.94509
MSE (mean squared error):	0.394509
R-Squared:	0.583207
Standard error:	0.628099
p-value (significance):	0.0038422

Individual trend lines:

Panes		**Line**		**Coefficients**				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Fitted Land Price (RM/ psf)	Distance to KL(km)	0.0038422	10	Distance to KL(km)	-0.0696839	0.0186286	-3.74069	0.0038422
				intercept	6.41878	0.612026	10.4878	< 0.0001

Figure 5.1.3. Details of Trend line for Land Price in 2017 against Distance to KL by city

5.2 Choice of Location

As mentioned earlier, the choice of satellite city proposed for CFLD is made based on the prospective profits by the end of the real estate project. In order to compare the prospective profits and ROIs across different satellite cities, the scope of the real estate project is declared by assuming a standardized construction period. In particular, it is assumed that CFLD would purchase the empty land in 2017, and the construction of the industrial park would be completed in 3 years. In other words, the newly-built industrial park would be available in the real estate market by 2020. The forecasted property prices in 2020 and fitted land prices in 2017 of the satellite cities under study are compared in Figure 5.2.

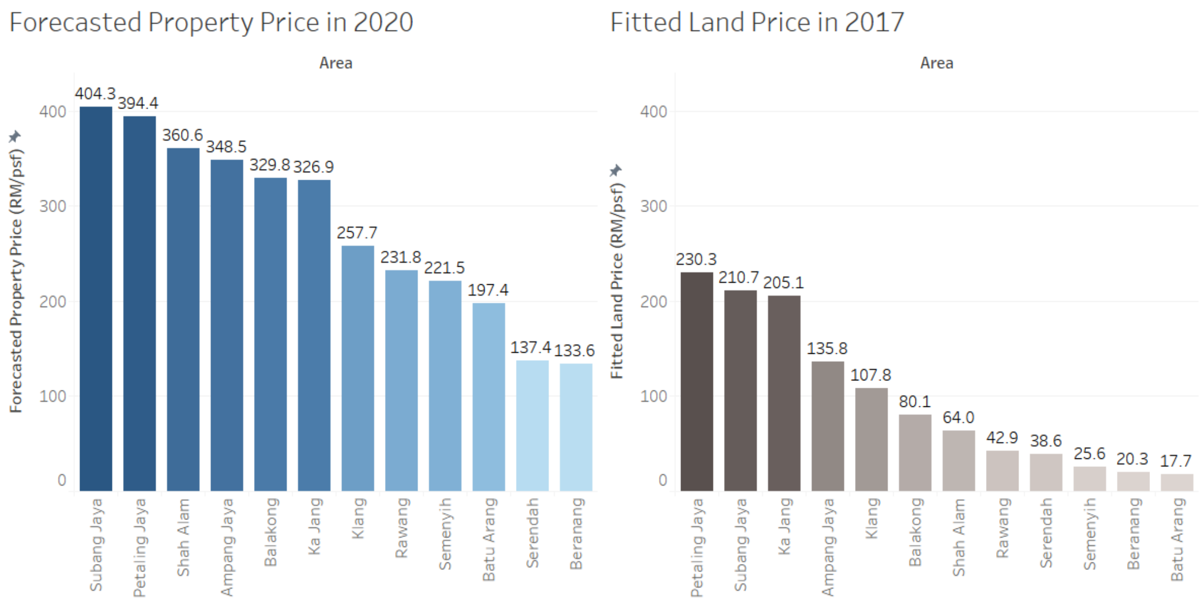


Figure 5.2 Comparison of forecasted property price and fitted land price

As shown in Figure 5.2, the huge gap between the estimated value of property in 2020 and the current land value indicates the potential investment opportunities. The optimal location of investment can be determined based on the predicted profits and annual ROIs listed in Table 5.2.

City/Town	Profit in 3 years (psf)	Annual ROI in 5 years (%)
Shah Alam	**296.6**	91.9
Balakong	**249.7**	58.2
Ampang Jaya	**212.7**	23.7
Semenyih	195.9	**158.6**
Subang Jaya	193.6	9.3
Rawang	189.0	86.9
Batu Arang	179.7	**214.0**
Petaling Jaya	164.0	4.7
Klang	149.9	19.8
Ka Jang	121.8	2.1
Beranang	113.3	**112.6**
Serendah	98.7	45.7

Table 5.2 Predicted profits and ROIs of the satellite cities

With the highest prospective profits of 296.6 RM/psf, **Shah Alam** is proposed as the optimal location of investment. Shah Alam is situated within the Petaling District and a small portion of the neighbouring Klang District. As the state capital of Selangor, it bears superior infrastructure and financial support provided by the government. The fastgrowing population in Shah Alam would provide sufficient labor supply to the newly built manufactoring plants or any other industrial property.

5.3 Investment Plan

With the advantaged geographical location and the stable economic environment, Shah Alam has attracted a number of real estate developers. The recent property development component consists of corporate towers, office suites, servied apartments, hotels and shopping mall (Priya Menon, 2016). In addition, many top multinational companies, including Schlumberger, Nestle, Pfizer, and Panasonic, set up their base in Shah Alam.

In general, there are three types of properties that real estate developers build, in particular, commercial properties, residential properties, and industrial properties. From the following heatmap of predicted prices of different types of properties, it can be observed that residential properties gives higher returns compared to industrial and commercial properties. Therefore, it is suggested that CFLD focus their investment on residential properties to gain more profits. The investment on industrial properties should align with the traditional key industries with local advantages, including electrical & electronics, transportation, life sciences, and manufacturing industry. It is believed that the influx of investors would contribute to the rapid economic growth and transform Shah Alam to a smart city in the future.

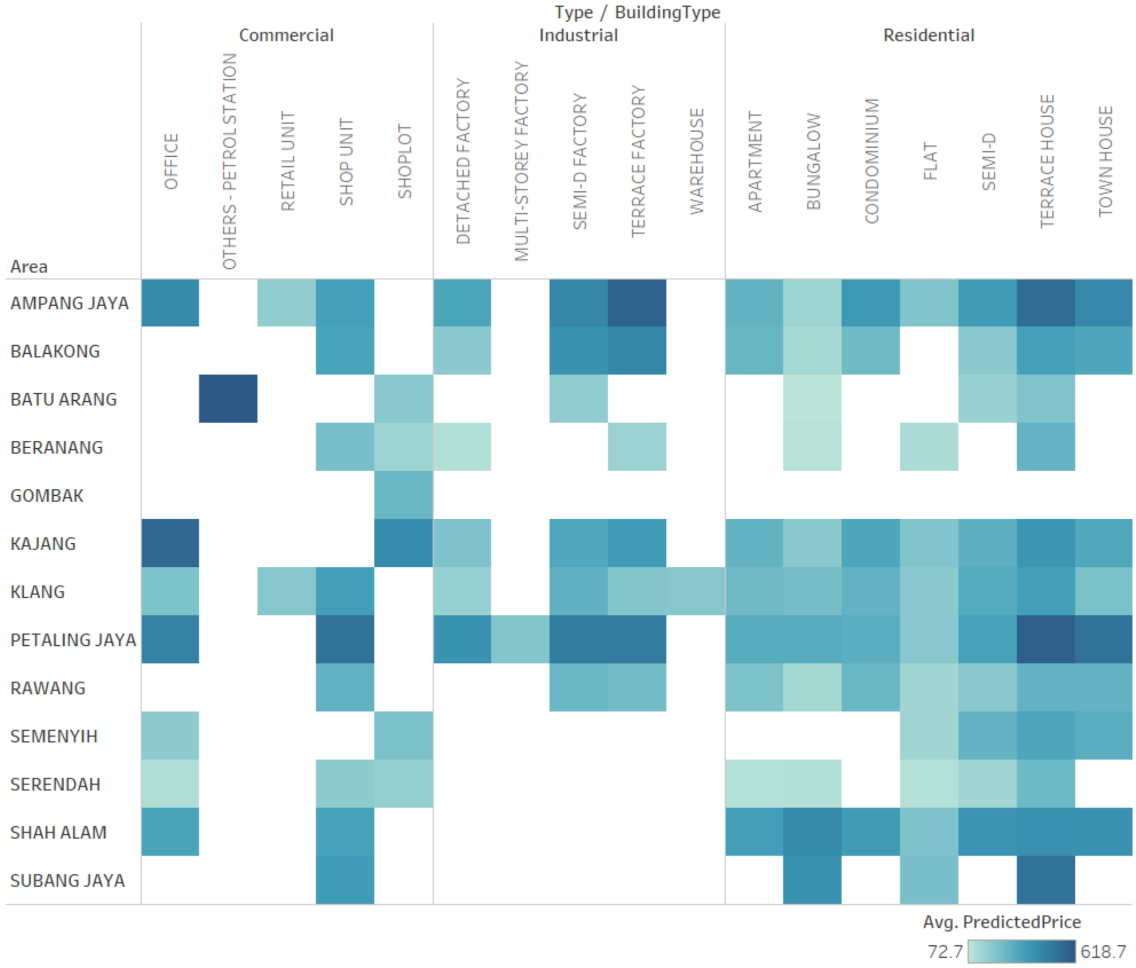


Figure 5.3 Heatmap of predicted prices of different building types

6. Conclusions and Further Work

6.1 Summary of key business insights: land choices and potential profits

The report provides an analysis and evaluation of land and property prices in 15 potential cities, which are 50km to Kuala Lumpur city centre, and provides the optimal investment strategy on one city based on the key analytic insights.

Methods of analysis include data sourcing, modelling, and evaluation. Innovative data sourcing methods, including scraping data from websites and making use of Google Map API to retrieve the nearest distance to amenities, were practiced along with traditional methods. Random forest regression model was performed to forecast the predicted property prices and empty land prices using predictors like demographics, economics, building type, and infrastructure. Using the predicted prices, profit and ROI were calculated to evaluate the best city to make investment.

The report concludes that Shah Alam is the best place to invest within next 3 years. It would result the highest profit of 296.6 Ringgit per square feet in additional to the fact that it has the established infrastructure for future development. Different types of properties, consisting of residential, industrial and commercial, were further analysed in Shah Alam state and found out that residential properties would provide higher return compare to the rest. The recommendations for expansion plan for Shah Alam was made and included industry focus on electric and electronic sector, transport equipment and life science.

6.2 Further Improvement on approaches of data sourcing and modelling

Many different adaptations, and improvements can be done on this report. Future work concerns more on data sourcing and modelling. Crime rate, natural disasters and living cost could not be collected at the town level due to lack of data although there were written news. Sentiment Analysis can be performed on written news to retrieve the above data. With those additional data, more precise evaluation can be done on land price prediction. In term of modelling, Monte Carlo simulation can be implemented to achieve probability distributions of possible outcome values. Probability distributions are realistic way of describing uncertainty in variables of a risk analysis.

7. Bibliography

[1] Menon, Priya. "Inevitable growth in Shah Alam." Metro News | The Star Online. February 14, 2016. https://www.thestar.com.my/metro/community/2016/02/15/inevitable-growth-in-shah-alam-selangors-capital-with-its-numerous-housing-developments-the-lrt-and/.

[2] Wu, C., Ye, X., Ren, F., Wan, Y., Ning, P., & Du, Q. (2016). Spatial and social media data analytics of housing prices in Shenzhen, China. PloS one, 11(10), e0164553.

8. Appendix

Appendix A – List of coefficients based on LASSO regression

```
(Intercept)                           3.616219e+02
(Intercept)                          .
AreaBALAKONG                         .
AreaBATU ARANG                      -6.907796e+01
AreaBERANANG                        -3.482757e+01
AreaGOMBAK                           .
AreaKAJANG                            8.601561e+00
AreaKLANG                           -1.717977e+01
AreaPETALING JAYA                    .
AreaRAWANG                           .
AreaSELAYANG                         .
AreaSEMENYIH                         .
AreaSERENDAH                        -7.339023e+01
AreaSHAH ALAM                        .
AreaSUBANG JAYA                      .
Year                                 .
TypeIndustrial                       .
TypeResidential                     -1.323997e+01
BuildingTypeBUNGALOW                -7.873000e+01
BuildingTypeCONDOMINIUM              .
BuildingTypeDETACHED FACTORY        -8.421976e+01
BuildingTypeFLAT                    -9.810692e+01
BuildingTypeMULTI-STOREY FACTORY    -1.726137e+02
BuildingTypeOFFICE                   .
BuildingTypeOTHERS - PETROL STATION  .
BuildingTypeRETAIL UNIT             -1.553101e+02
BuildingTypeSEMI-D                   .
BuildingTypeSEMI-D FACTORY            4.028393e+01
BuildingTypeSHOP UNIT                 1.016011e+02
BuildingTypeSHOPLOT                   1.459038e+01
BuildingTypeTERRACE FACTORY           2.200838e+01
BuildingTypeTERRACE HOUSE             1.070797e+02
BuildingTypeTOWN HOUSE                4.548525e+01
BuildingTypeWAREHOUSE                .
LandArea                             .
StateSelengor                        .
CityArea.km2.                        .
City.Town.VillageTown               -1.083717e+02
Population                           .
CPI                                  .
Pop.Density                           1.626966e-02
dis_Klcenter_km                      .
time_Klcenter_min                    .
dis_airport_km                      -1.619692e-01
time_airport_min                     .
min_dist_hospital_2km                .
min_dist_school_2km                 -2.712899e-03
dist_conv_store_2km                 -5.190132e-03
dist_train_10km                       2.910791e-04
dist_min_police_2km                  .
```

```
dist_bus_500m                       -5.373235e-02
no_hospital_2km                      .           
no_train_station_10km                .           
no_conv_store_2km                    .           
no_manufacturing_1km                 6.556059e-01
no_manufacturing_500                 .           
no_police_2km                        .           
no_bus_station_500m                  1.893201e+00
no_school_1km                        .           
```

Appendix B - List of 27 predictors used for Random forest model

Predictors		
Area within Greater KL	Year	Building Type
Land Area Size (km2)	City Area (km2)	State
City/Town/Village	Population Density	Consumer Price Index
Distance to KL city center	Travelling time to KL city center	Distance to KL International Airport
Travelling time to KL International Airport	Minimum distance to hospital/clinics	Minimum distance to Educational Institutions
Minimum distance to Shops	Minimum distance to Train Station	Minimum distance to Bus stop
Minimum distance to Police Station	Number of hospital/clinics within 2km radius	Number of Educational Institutions within 1km radius
Number of shops within 2km radius	Number of Train Stations within 2km radius	Number of Bus Stops within 500m radius
Number of Police Station within 2km radius	Number of Manufacturing sites within 500m radius	Number of Manufacturing sites within 1km radius