Predicting Singapore Private Property Prices

Prepared by
Vishali NALLA
SUN Shuangtian
TAN Yong Ying
TONG Wen Liang Samuel

21st November 2018

Contents

1. Abstract

The housing demand in Singapore has grown significantly over the years, in proportion to the increase in population density. In this study, we have applied various machine learning techniques to help multiple stakeholders to identify features that have the most significant impact on property prices. Consequently, stakeholders such as property developers will be able to capitalise on the knowledge of specific features that are important in driving property prices and to test these assumptions prior to incurring project cost. Each supervised machine learning model was built based on six two-tiered domains. Each domain consist of a particular Type of Sale corresponding to a particular Property Type. Among all models, Extreme Gradient Boosting Regressor achieved the highest R^2 score across all domains. We have also identified the most important features overall across all domains, and common important features across each type of sale - Resale and New Sale. Finally, the impact of these features were visualized and interpreted to generate interesting insights.

2. Objective and Use Case

With a population of over 5.5 million in an area of 721.5 km^2, the island state of Singapore is the third most densely populated country in the world at 7,796 persons/km^2 in 2017. The significant increase in population density thus translate to the need for greater housing demand. The ability to determine non-policies factors that would contribute to the prices of private residential housing in Singapore would be useful to multiple stakeholders. Both property developers and potential private home owners generally lack insights on the most important features that affect property prices. Therefore, various machine learning models will be implemented to predict property prices and to highlight the most important features for each property type and type of sale.

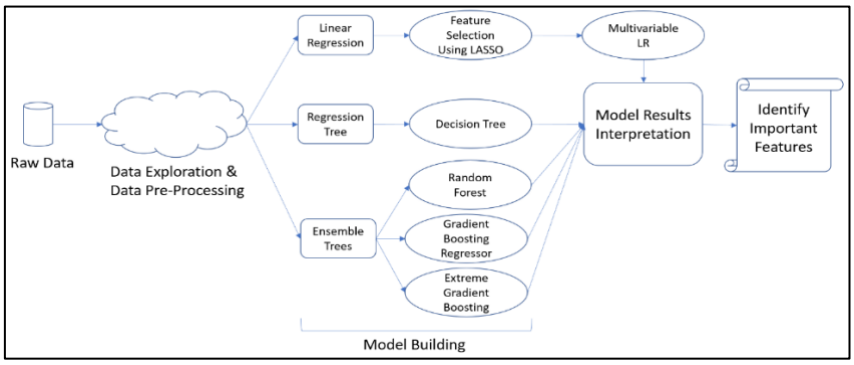


Figure 1: Project Overview

3. Data Exploration & Data Pre-Processing

3.1 Dataset Description

This dataset contains the transaction information of Singapore private properties including apartment, condominium and executive condominium from January 2013 to January 2016. The data set has over 50,000 rows with over 70 different features organised into four main categories. In this regression problem, we will be predicting the unit price of the properties per square feet i.e, *Unit Price ($psf)* [Appendix I].

Original REALIS	
SN	Feature Name
1	Project Name
2	Address
3	No. of Units
.	.
.	.
18	Area (sqm)
19	Property Type
20	Type of Sale

Facilities	
SN	Feature Name
1	Security
2	Carpark
3	Entertainment
.	.
.	.
6	Family-Friendly
7	Fitness
8	Function Room

Distance	
SN	Feature Name
1	Distance to CBD
2	Distance to nearest childcare centre
.	.
.	.
21	Distance to nearest hawker centre and market
22	Distance to nearest bus interchange

Demographic	
SN	Feature Name
1	No of Residents (Million)
2	Age 65 and above (% of total residents)
.	.
.	.
6	Condominiums and Private Flats Households (% of total households)
7	University Qualification (% of total residents)

Figure 2: Dataset Description Summary

3.2 Basic Data Cleaning

Prior to model building, standard data cleaning techniques were applied to handle common issues with raw data such as the presence of missing data and outliers. More details on the basic data cleaning techniques employed can be seen in Appendix II.

3.3 Feature Constructions and Removal

Business domain knowledge was applied to understand the features' meaning and relevance to the regression problem. Thereafter, new features deemed to be important in building our models were created beyond the existing dataset. Conversely, features that were deemed to be unnecessary or irrelevant to our problem were removed. While part of this analysis may be subjective based on intuition, it is important to apply contextual knowledge to ensure that the models take into account of any feature that might be important in predicting property prices.

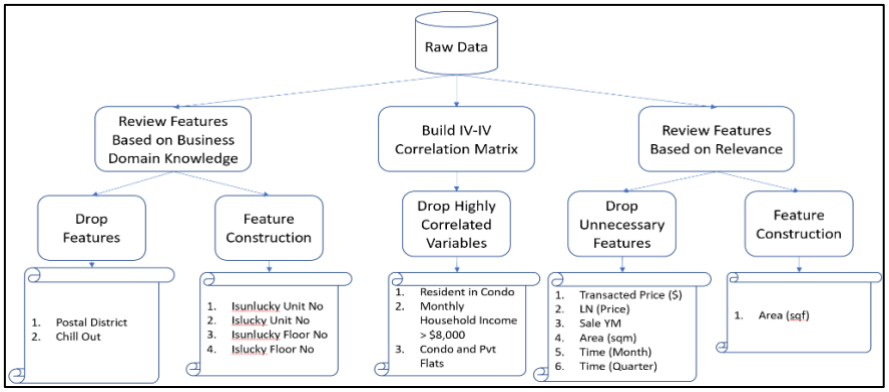


Figure 3: Feature Construction & Removal Overview

3.3.1 Based on Domain Knowledge

Drop Features - During the data pre-processing step, it is common to find features that have identical or overlapping meaning. In this case, the relevance and meaning of feature-pairs *Postal District* & *Planning Area* and *Chill Out* & *Food/Dining* overlap. Therefore, *Postal District* and *Chill Out* were removed from the dataset.

Feature Construction - Based on social-cultural factors in the local context, the following features were created as they are likely to be important in influencing buyers' decision in purchasing a property:

Newly Constructed Feature	Description
Islucky Unit No	Lucky numbers include 8, 88, or 888
Islucky Floor No	Lucky numbers include 8 or 88
Islunlucky Unit No	Unlucky numbers include 4, 44, or 444
Isunlucky Floor No	Unlucky numbers include 4, or 44

Table 3.1: Auspicious and Inauspicious Floor/Unit Features

In Singapore, auspicious or inauspicious numbers are typically considered when one purchases new goods or services. This belief could therefore extend to the purchase of properties as well.

3.3.2 Based on Multicollinearity

The removal of features have been subjective and based on intuitive business domain knowledge thus far. To further support our data pre-processing step, we employed a more objective analysis through the introduction of an Independent Variable – Independent Variable correlation matrix to identify highly correlated features. Objectively, we want to remove any feature in a pair of features that show a strong correlation to prevent multicollinearity. More details on the removal of highly correlated features can be seen in Appendix III.

3.3.3. Based on Relevance

Drop Features - Additionally, features that were not relevant or useful have to be removed from the dataset to ensure robustness of the models. For example, feature such as *Transacted Price ($)* would not be useful since it overlaps with our target variable *Unit Price ($psf)*.

Feature Construction - Since our target variable *Unit Price ($psf)*, it would be more useful to convert the existing *Area (sqm)* feature into *Area (sqf)* to make our models more robust.

3.4 Feature Conversion

It is imperative to take steps to ensure that the features are appropriately transformed to ensure compliance with the statistical assumption of the machine learning techniques employed.

Log Distance - Explicitly for distance to nearest amenities features, based on business intuition it can be assumed that short and long distances are more important in predicting property prices. Therefore, we want to perform data transformation on these features to reflect this intuition. Additionally, when we visualize an example of one of the distance features, *bus stop km*, onto a histogram in Figure 4, we can see that the datapoints are typically skewed. Correspondingly, the data distribution does not closely follow the diagonal line in the probability plot that represents normal distribution.

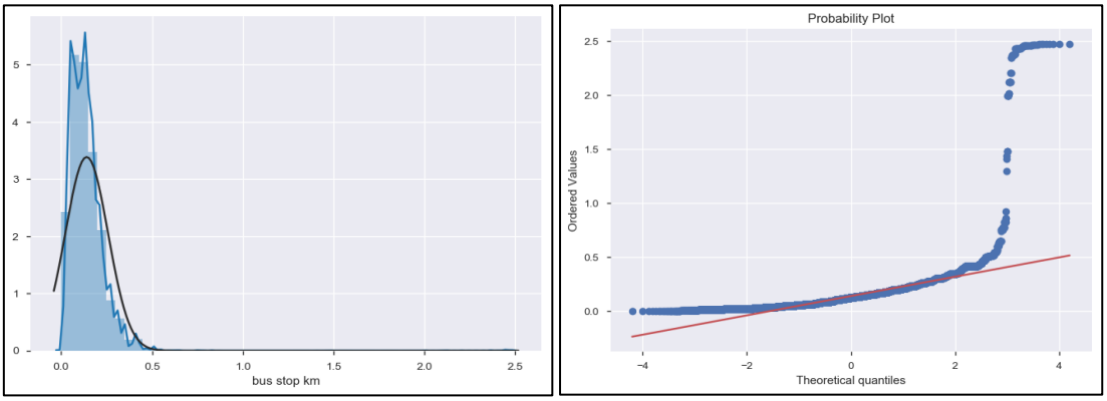


Figure 4: Histogram and Probability Plot of "bus stop km" prior to log transformation

Upon log transformation, the data distribution now follows a normal distribution more closely.

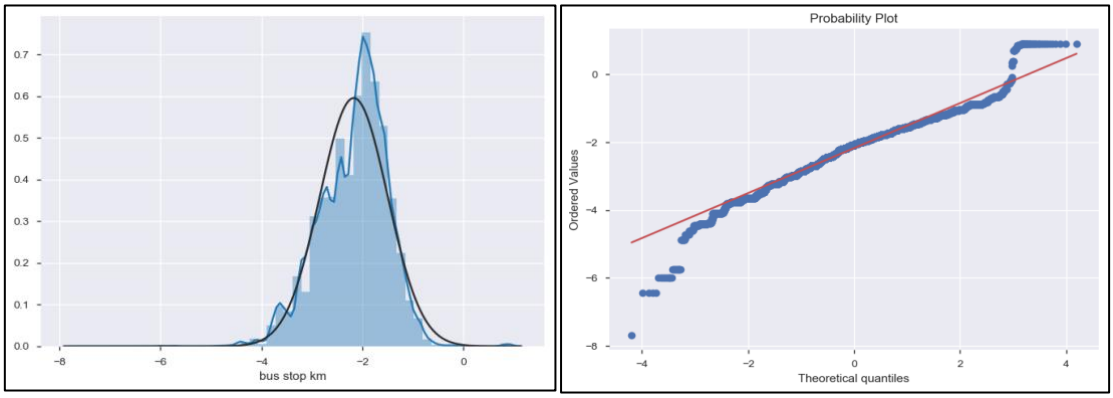


Figure 5: Histogram and Probability Plot of "bus stop km" after log transformation

Feature Conversion for LR Models - Since Linear Regression has greater statistical limitations imposed on its features, the following data transformation were performed. First, categorical variables were re-coded as dummy variables by one-hot encoding (n columns for n levels). Second, all variables were scaled using sklearn.preprocessing.MinMaxScaler. This means that all features had values in the range of [0,1]. If features were not normalized, it would be unfair to evaluate feature importance using the coefficients since their magnitude would be highly influenced by extremely large or small values.

4. Model Building Methodology

To ensure that the models are robust in predicting property prices, the models were built based on six two-tiered domains corresponding to each property type and type of sales. From Figure 6, there is a distinct difference in unit price for each property type and type of sale.

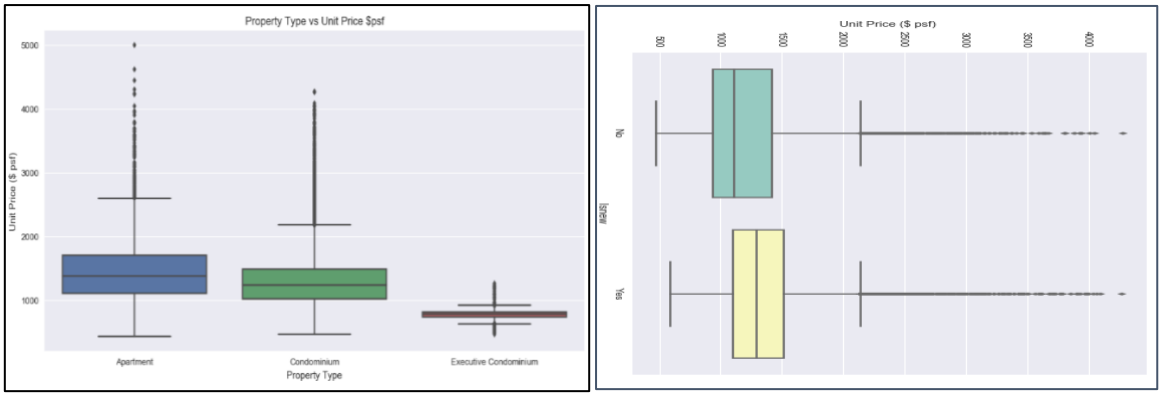


Figure 6: (a) Left - Boxplot of Unit Price ($psf) per Property Type (b) Right - Boxplot of Unit Price ($psf) per Type of Sale

Therefore, the model building methodology can be summarized based on Figure 7 below:

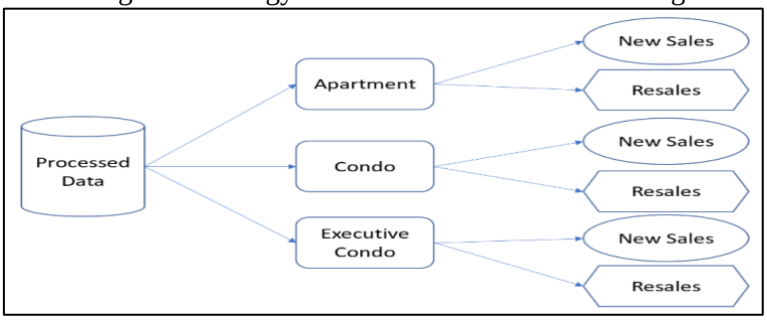


Figure 7: Model Building Methodology Overview

5. Model Results Interpretation

5.1 Results

Domain	Results	Multivariable LR (After Feature Selection with LASSO)	Decision Tree	Random Forest Regressor	Gradient Boosting Regressor (GBR)	Extreme Gradient Boosting Regressor (XGBRegressor)
Apartment New Sales	**R^2**	0.872	0.946	0.954	0.961	0.966
	MSE	25,369.94	10,763.35	9,032.06	7,628.89	6,744.26
Apartment Resales	**R^2**	0.808	0.876	0.920	0.922	0.928
	MSE	52,413.64	33,864.23	21,899.48	21,300.09	19,681.25
Condominium New Sales	**R^2**	0.850	0.952	0.964	0.967	0.967
	MSE	19,323.00	6,204.14	4,658.75	4,248.71	4,180.87
Condominium Resales	**R^2**	0.796	0.864	0.922	0.932	0.932
	MSE	35,055.74	23,359.82	13,374.75	11,659.23	11,610.83
Executive Condominium New Sales	**R^2**	0.572	0.728	0.788	0.802	0.803
	MSE	1,316.34	836.21	651.39	609.41	605.93
Executive Condominium Resales	**R^2**	0.833	0.825	0.872	0.887	0.889
	MSE	1,824.53	1,915.71	1,397.56	1,236.05	1,213.52

Table 5.1: Predictive Model Regression Results

Out of the predictive models employed for each domain, we can deduce that XGBRegressor is the best performing model where it achieved the highest R^2 score and lowest MSE score for all domains. To determine how well the models fit the data, we need to take a statistical measure of how close each data point fits the regression line. By taking XGBRegressor from the Apartment – New Sales and Resales domains as an example, we can visualize its goodness-of-fit as shown in the diagrams below:

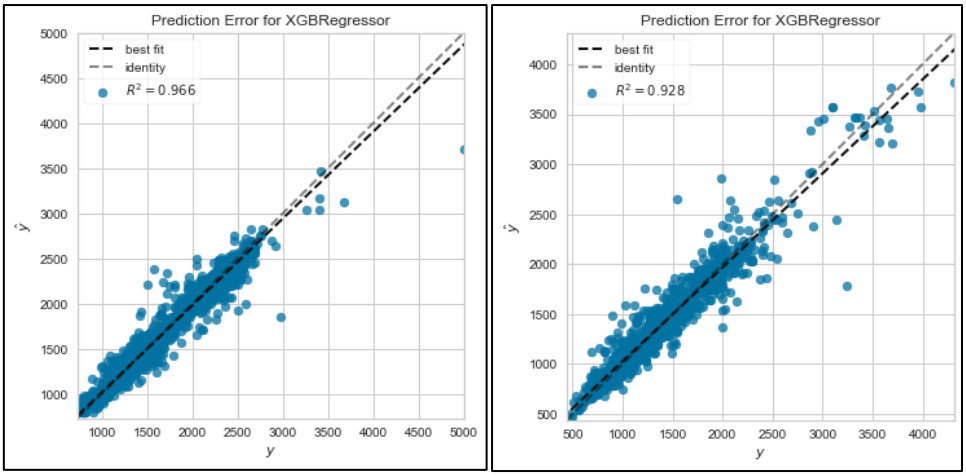


Figure 8: Prediction Error Plot for XGBRegressor on Apartment: (a) New Sales and (b) Resales

In general, the higher the R^2, the better the model fits the data. However, the measurement of R^2 alone is not sufficient to determine biasness of the predictions. To do so, we employed the use of residual plots to analyse the variance of error. A random dispersion of data points around the horizontal axis tells us that our model is performing well. Additionally, from the histogram on the right, we can also see that the error is normally distributed around zero, providing another indication of a well fitted model.

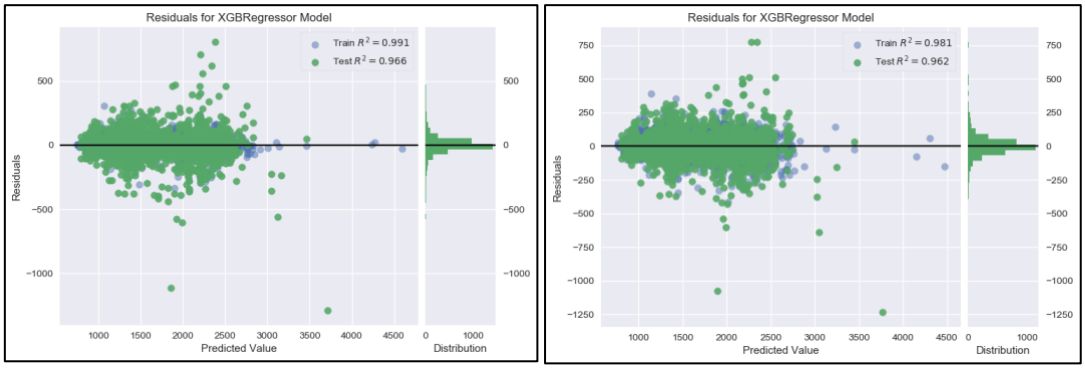


Figure 9: Residual Plot for XGBRegressor on Apartment: (a) New Sales and (b) Resales

5.2 Model Comparison

5.2.1 Linear Regression:

Feature selection was done using lasso regression before the selected features were used to build a basic linear regression model. Lasso regression is mainly used to prevent the model overfitting to the data through the L1-norm regularization parameter, but it also performs feature selection as it sets the coefficients of non-important features to zero, which effectively drops the features from the regression model. Lasso regression was performed with five values of the alpha parameter: 0.01, 0.1, 1, 10 and 100 and for each alpha value, the model was optimized using the objective function:

$$SS_{res} \text{ (Residual sum of squares)} + [\alpha * \text{(sum of absolute value of coefficients)}]$$

Across all property and sale types, α of 0.01 gives the lowest total cost. However, it also resulted in a high number of features being selected (e.g. 78 out of 94 features had non-zero coefficients for new apartment sales data subset when α = 0.01). By plotting the total cost across the five α values, we see that α of 1 still gives a low enough total cost. Number of selected features dropped significantly from 78 to 48. Therefore, we used α = 1 for all the six data subsets to select the features for the basic linear regression model. Ridge regression was not considered, because the fact that an α value close to 0 gives us the lowest total cost means that the basic linear regression model without regularization is the best regression model.

5.2.2 Decision Tree

Compared to linear regression, decision trees do a better job at capturing the non-linearity in the data by dividing the space into smaller sub-spaces.

5.2.3 Ensemble Trees

Random Forest Regressor - The hypermeters of the Random Forest Regressor is almost identical to that of the Regressor Tree. However, it differs by the addition of more randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This diversity in features thus makes the model better for feature selection than the Decision Tree.

Gradient Boosting Regressor (GBR) - GBR builds its trees in a sequential manner such that it takes into account of the error from the previously trained trees. Each new tree thus helps to correct the error made by the previous tree iteratively to better train the model.

Extreme Gradient Boosting Regressor (XGBRegressor) - Both XGBRegressor and GBR follows the principle of gradient boosting. However, XGBRegressor differs in terms of its modelling details where it uses a more regularized model formalization to control over-fitting. Therefore, this gives the XGBRegressor an edge over the GBR.

6. Feature Importance

In XGBRegressor, the evaluation of important features is performed based on *feature_importances_*, which allows us to gain a sense of which features have the most effect on the predictability of the Unit Price ($ psf). Feature importance is derived based on the frequency of which a feature is used to split the tree. Therefore, this section will be divided into three parts, where we will be discussing on the common important features selected by the XGBRegressor overall for all domains, common important features across type of sales and other interesting insights based on feature importance.

6.1 Overall Important Features

Domain	Rank	Feature Name	Score	Domain	Rank	Feature Name	Score
Apartment New Sales	**1**	Area (sqf)	0.358	**Apartment Resales**	**1**	Area (sqf)	0.176
	2	Floor No (Final)	0.096		**2**	Age	0.118
	3	Childcare centre km	0.027		**3**	Floor No (Final)	0.068
Condominium New Sales	**1**	Area (sqf)	0.372	**Condominium Resales**	**1**	Area (sqf)	0.201
	2	Floor No (Final)	0.104		**2**	Age	0.098
	3	Junior college km	0.025		**3**	Floor No (Final)	0.060
Executive Condominium New Sales	**1**	Area (sqf)	0.396	**Executive Condominium Resales**	**1**	Area (sqf)	0.188
	2	Floor No (Final)	0.098		**2**	Age	0.177
	3	CBD km	0.040		**3**	Floor No (Final)	0.107

Table 6.1: Top Three Important Features across All Domains

The top three most important features across all domains are indicated in table 6.1 above. From the table, we can see that both *Area (sqf)* and *Floor No (Final)* consistently appear throughout with a relatively high importance score. However, in order to truly understand how these features affect our target variable, we will visualize these relationships to gain further insights.

6.1.1 Area (sqf)

Figure 10 below shows the regression plot of our target variable *Unit Price ($psf)* against *Area (sqf)*. In general, the plots show that as the size of the property increases, the *Unit Price ($psf)* decreases. However, we note that the only exception to this trend comes from the Condominium Resales domain, where the trend is inversed that of the other domains. While there could be multiple factors leading to such observations, it is interesting to consider certain specific reasons that could have led to this observation. For example, based on business intuition, Condominiums differ from the other property type as some Condominium units are penthouses, where the area of such property units are significantly much larger than other units. Additionally, given that the observation is based on Resales, where properties in this category are typically of an older age, it is interesting to note that over the years, the size of each Condominium units are decreasing due to the land space constraint in Singapore. Therefore, older Condominium units are typically larger than recently constructed Condominium units.

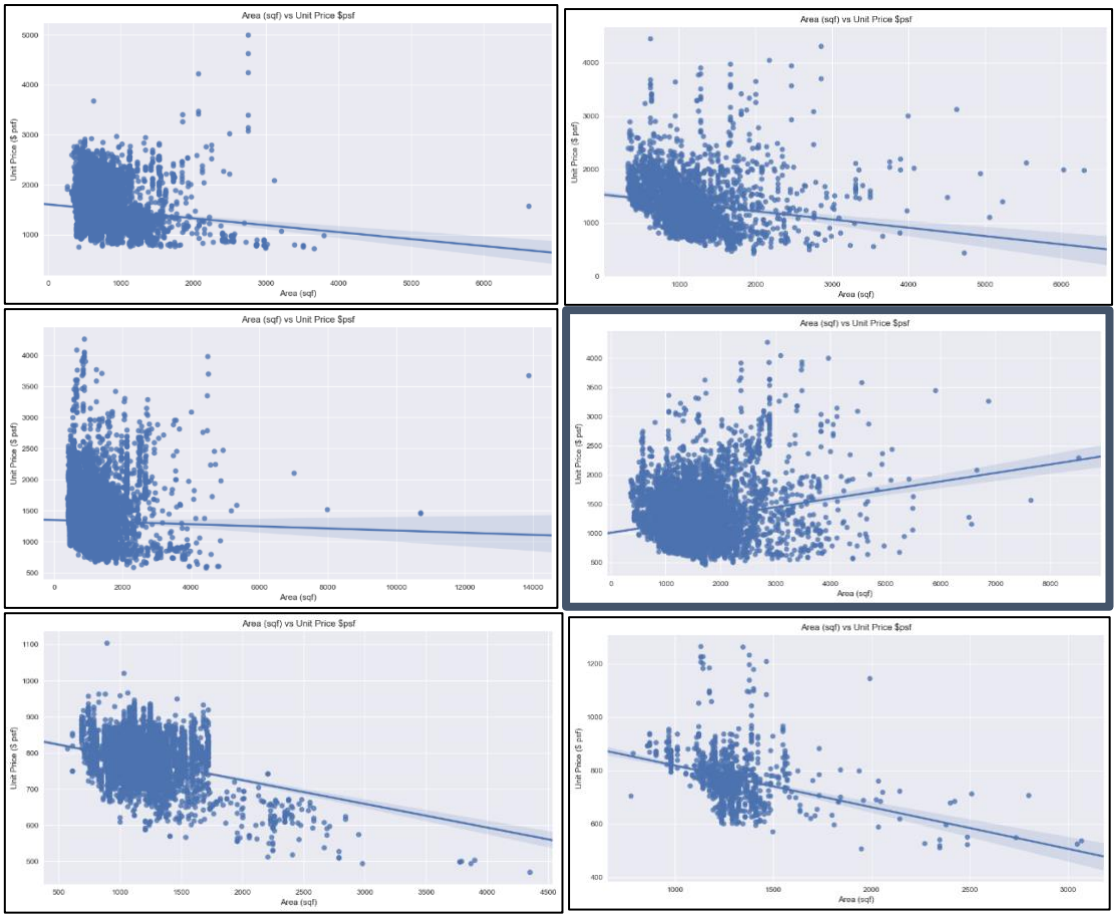
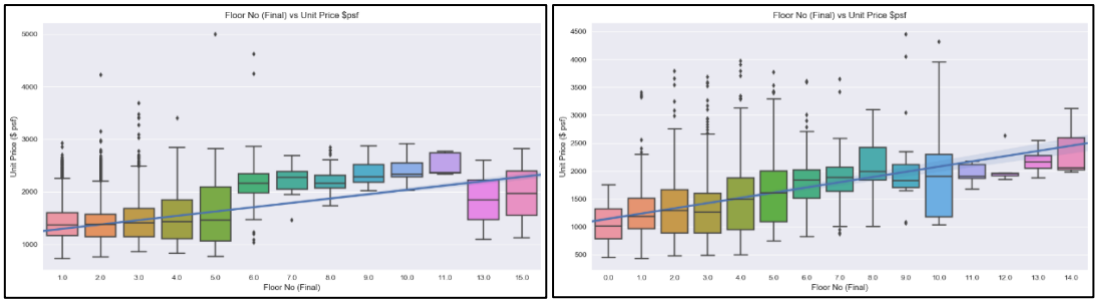


Figure 10: Regression Plot of Unit Price ($psf) against Area (sqf) for (a) Top Left–Apartment New Sales; (b) Top Right–Apartment Resales; (c) Middle Left–Condo New Sales; (d) Middle Right–Condo Resales; (e) Bottom Left–Executive Condo New Sales; (f) Bottom Right–Executive Condo Resales

6.1.2 Floor No (Final)

The next most important variable that consistently appears in the top three most important variable list across all domains is *Floor No (Final)*. Based on Figure 11 below, it is clear that as floor number increases, the *Unit Price ($psf)* increases as well. However, we can also see that the regression plot for Condominium New Sales and Resales and Executive Condominium Resales differ from the other property type, where we can see a spike in *Unit Price ($psf)* after a certain floor number. Similar to our discussion in section 6.1.1 above, based on business domain knowledge, this spike in *Unit Price ($psf)* could be due to the presence of significantly larger penthouses in Condominiums. For Executive Condominiums, prior to 2013, when the Singapore Government implemented measures to curb the size of Executive Condominium units to 1,722 sqf, Executive Condominiums have penthouses too. Therefore, this could be reflected in Figure 11 below, where Executive Condominium Resales, which are typically properties of an older age, see a spike in *Unit Price (psf)* after a certain floor number.



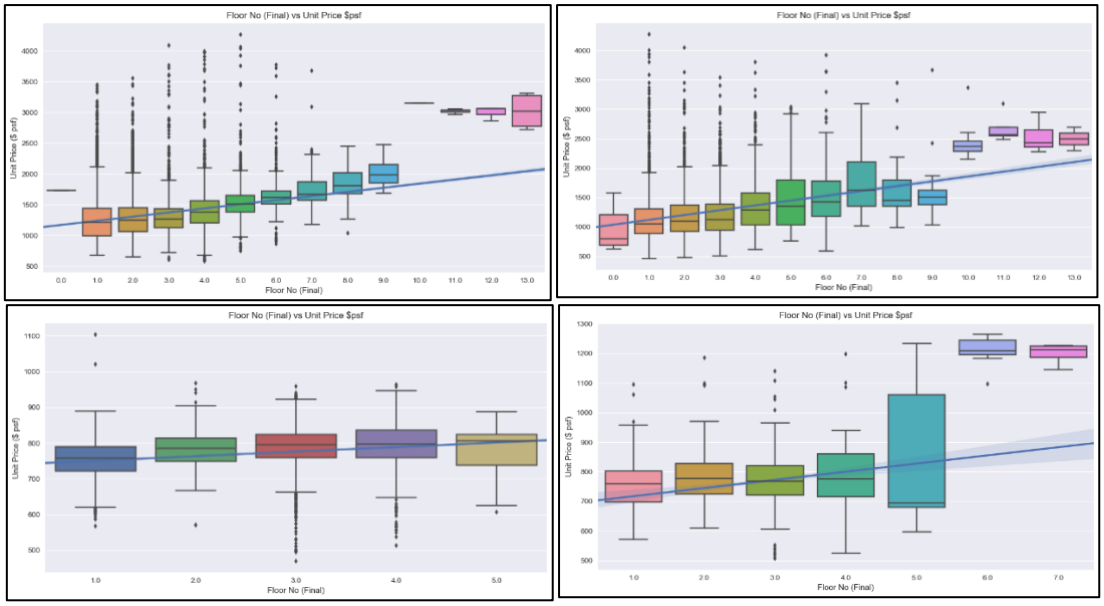


Figure 11: Regression Plot of Unit Price ($psf) against Area (sqf) for (a) Top Left–Apartment New Sales; (b) Top Right–Apartment Resales; (c) Middle Left–Condo New Sales; (d) Middle Right–Condo Resales; (e) Bottom Left–Executive Condo New Sales; (f) Bottom Right–Executive Condo Resales

6.1.3 Distance to Public Transportation

Even though the features representing distance to public transportation did not appear in the top three most important features across all six domains, it is important to note that these features do appear to be common across all six domains among the top ten most important features. More details of the top ten most important features across all domains can be found in Appendix IV. Based on Figure 12 on the right, we can see that *Unit Price ($psf)* is inversely related to two features representing distance to public transport – *Bus Distance (log)* and *MRT Distance (log)*. As the distance to bus stop and to MRT station decreases, the *Unit Price ($psf)* increases.

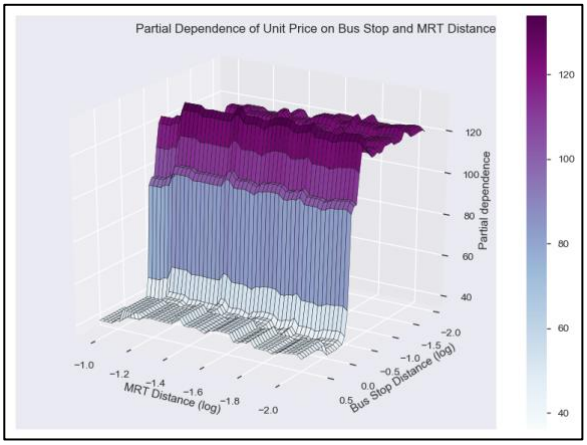


Figure 12: Partial Dependence Plot of Unit Price ($psf)against distance to MRT and distance to bus stop

6.2 Important Features across Type of Sale

Having visualized all important common features across all six domains, we want to identify features that are common across each type of sale to obtain further insights on whether there are features that are only unique to each particular sales type.

6.2.1 Resales

Figure 13 shows the top ten most important features for the resale of all property types. Based on the figure, we can see that *Age* appears to be a significantly important feature that consistently appear across all property types as the second most important feature.

Apartment

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	CBD km
2	Age	7	bus stop km
3	Floor No (Final)	8	exercise facility km
4	elderly care centre km	9	community club km
5	park km	10	kindergarten km

Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	MRT station km
2	Age	7	exercise facility km
3	Floor No (Final)	8	No of Units
4	CBD km	9	secondary school km
5	bus stop km	10	community club km

Executive Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	CBD km
2	Age	7	elderly care centre km
3	Floor No (Final)	8	bus stop km
4	kindergarten km	9	exercise facility km
5	community club km	10	Uni Qualification (% of total)

Figure 13: Common Important Features across Resales Based on XGBRegressor

Based on Figure 14, we can see that as *Age* is inversely related to *Unit Price ($psf)*. As *Age* increases, the *Unit Price ($psf)* decreases. However, it is also interesting to note that for Apartment and Condominiums, we can see a spike in *Unit Price ($psf)* at around age 24 for Apartments and age 23 to 26 for Condominiums. Based on business domain knowledge, the sudden rise in price could be due to demand driven by property investors who are interested in profiting from a potential en bloc sale, which typically occurs for properties that were more than 20 years old at the point of sale.

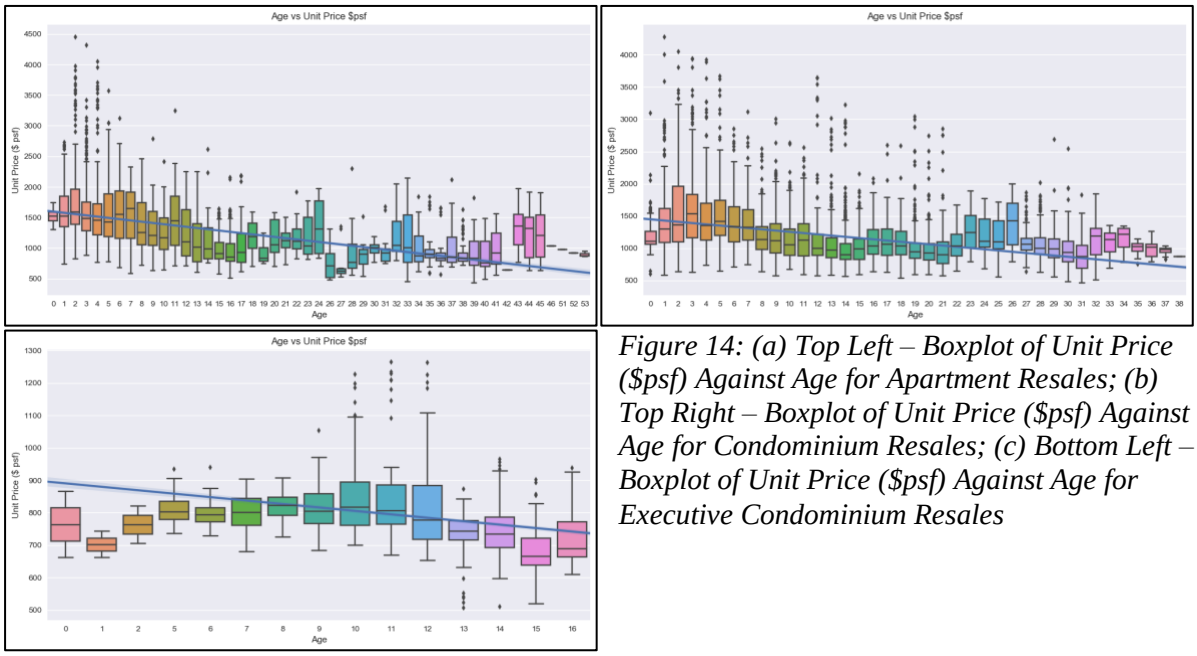


Figure 14: (a) Top Left – Boxplot of Unit Price ($psf) Against Age for Apartment Resales; (b) Top Right – Boxplot of Unit Price ($psf) Against Age for Condominium Resales; (c) Bottom Left – Boxplot of Unit Price ($psf) Against Age for Executive Condominium Resales

Additionally, two other features that are common across all property types for Resales are *community club km* and *exercise facility km*. It is possible that the age group of buyers of resale properties are older, and the distance to such amenities are factors in their decision to purchase a property.

6.2.2 New Sales

As we proceed to analyse the important features for the new sales of all property types, Figure 15 shows that the most common important features that appears are *childcare centre km* and *kindergarten km*. This shows that the home buyers of new properties could possibly be new young families that are buying their first home, and distance to childcare service amenities are an important factor in their decision to buy a new home.

Apartment

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	shopping mall km
2	Floor No (Final)	7	supermarket/hypermarket km
3	childcare centre km	8	junior college km
4	kindergarten km	9	elderly care centre km
5	bus stop km	10	Age

Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	top primary school km
2	Floor No (Final)	7	No of Units
3	junior college km	8	primary school km
4	CBD km	9	childcare centre km
5	kindergarten km	10	MRT station km

Executive Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	community club km
2	Floor No (Final)	7	park km
3	CBD km	8	Isunlucky Floor No
4	exercise facility km	9	childcare centre km
5	URA growth area km	10	bus stop km

Figure 15: Common Important Features across New Sales Based on XGBRegressor

6.3 Other Interesting Insights

6.3.1 Auspicious and Inauspicious Numbers

In this section, we want to discuss other interesting features detected by the GBR. Even though GBR was not the best performing model, its performance came close to that of the XGBRegressor at most times. When we explored feature importance of GBR, auspicious/inauspicious features discussed in Data Pre-Processing step *Table 3.1* were identified as important features for New Sales of all domains.

Apartment

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	Islucky Unit No
2	Floor No (Final)	7	Isunlucky Unit No
3	Purchaser Address Indicator_Private	8	Islucky Floor No
4	Purchaser Address Indicator_HDB	9	junior college km
5	Isunlucky Floor No	10	hospital/polyclinic km

Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	Purchaser Address Indicator_N.A
2	Floor No (Final)	7	junior college km
3	Islucky Unit No	8	kindergarten km
4	Purchaser Address Indicator_HDB	9	MRT station km
5	Purchaser Address Indicator_Private	10	Age

Executive Condo

Rank	Feature Name	Rank	Feature Name
1	Area (sqf)	6	bus stop km
2	Floor No (Final)	7	Purchaser Address Indicator_HDB
3	Isunlucky Floor No	8	Purchaser Address Indicator_Private
4	Islucky Unit No	9	Purchaser Address Indicator_N.A
5	Islucky Floor No	10	URA growth area km

Figure 16: Common Important Features across New Sales Based on GBR

It is interesting to note that auspicious and inauspicious numbers are deemed to be more important for New Sales than Resales, since they have consistently appeared in the top ten most important features in New Sales but not in Resales across all property types.

7. Conclusion & Future Work

Private Residential Housing in Singapore is classified into three main property types which has huge variance along budget, area, lifestyle and facilities. To testify this fact, In our study we have applied various supervised machine learning models on six two-tiered domains to predict the price of the house and to identify important features dominating the price of the property. Extreme Gradient Boosting yielded highest R squared (0.80 - 0.96) across all domains. Relatively, new sales domain has better results than resales due to the limitation of fewer data points for resales. Interestingly, the important features are marginally different for Random Forest Regressor, Decision Tree, Linear Regression when compared to XGB/GBR. Leveraging on the features identified for the best model for each domain, Area (psf) and Floor No topped the list in all domains whereas subsequent important

features varied for each domain. Age, distance to community club and exercise centre are important for Resale properties whereas distance to childcare centre and kindergarten, auspicious/inauspicious numbers are important for new sale properties indicating that young families prefer buying new properties. In future, we plan to incorporate economic features such as GDP, housing policies, rental yield and key facilities of number of bedrooms, bathrooms etc to better predict the housing prices and to relatively measure the feature importance across different domains. By augmenting transactional data with more recent years housing data and eliminating extreme irregularities like very high unit priced transactions, we can further improve model predictions.

8. References

PHANG, S.Y. (2001). Housing Policy, Wealth Formation and the Singapore Economy. *Housing Studies*.
16, (4), 443-459. Research Collection School Of Economics.

Data.gov.sg. Private Residential Property Price Index. Retrieved from https://data.gov.sg/dataset/private-residential-property-price-index-by-type-of-property?resource_id=e779eb53-a1c0-4670-a5ac-43b5b534496d

Data.gov.sg. Indicators on Population, Annual. Retrieved from https://data.gov.sg/dataset/indicators-on-population-annual?resource_id=19362aee-440d-4f41-97af-adb9c212412d

Ray, S. (2017). Essentials of Machine Learning Algorithms. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

John, R. (n.d.). Simple Housing Price Prediction Using Neural Networks with TensorFlow. Retrieved from https://medium.com/@robertjohn_15390/simple-housing-price-prediction-using-neural-networks-with-tensorflow-8b486d3db3ca

Shalizi, C. (2006). Lecture 10: Regression Trees. Retrieved from http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Chen, J.H., Ong, C.F., Zheng, L. & Hsu, S.C. (2016). Forecasting Spatial Dynamics of the Housing Market Using Support Vector Machine. *International Journal of Strategic Property Management*. 21:3, 273-283.

Ng, M. (2013). No more massive EC with new cap on maximum size. Retrieved from https://www.stproperty.sg/articles-property/condominium/no-more-massive-ec-units-with-new-cap-on-maximum-size/a/100769

Choo, C. (2018). 13 of 40 properties sold en bloc in last three years were less than 30 years old. Retrieved from https://www.todayonline.com/singapore/13-40-properties-sold-en-bloc-last-three-years-were-less-30-years-old

Appendix I - Dataset Description

S No	Name of Features	Description
1	SN	Unique transaction ID
2	Project Name	The name of the project
3	Address	The full address of the property
4	No of Units	Number of units in the transaction
5	Area (sqm)	The area of the property in square meters
6	Type of Area	Strata or Land
7	Transacted Price ($)	The transacted price of the property
8	LN (Price)	LN of Transacted Price ($)
9	Nett Price ($)	Net transaction price
10	Unit Price ($ psm)	Transacted price divide by area in sqm
11	Unit Price ($ psf)	Transacted price divide by area in sqf
12	Sale Date	The sale date of the property
13	Sale YM	Year – Month of Sale Date
14	Time (Quarter)	Time lag by Quarter
15	Time (Month)	Time lag by month
16	Property Type	Condo, Apartment, Executive Condo
17	Tenure	Tenure of the Property
18	Completion Date	The completion date of the project
19	Type of Sale	New sale, sub sale or resale
20	Purchaser Address Indicator	If Buyer's address is private property or HDB
21	Postal District	Postal sector of the property
22	Postal Sector	Postal sector of the property
23	Postal Code	Postal code of the property
24	Planning Region	Planning region the property belongs to
25	Planning Area	Planning area the property belongs to
26	Address 1	Address of the project (duplicated variable)
27	Floor No	Floor of property
28	Floor No (Final)	Consecutive 5 floors binned into 1 category eg – Floor No 1-5 is 1 for Floor No (Final)
29	Unit No	Unit No of the transacted property
30	Age	Age of the transacted property
31	Tenure (New)	Derived from Tenure
32	Tenure (Final)	Derived from Tenure (New) ; 2 if Freehold, 1 otherwise
33	Lease Start Date	Start Date of Lease Period of property
34	Address for Geocode	Address of property derived from Google Map Geocoding API
35	Latitude	Latitude of the property
36	Longitude	Longitude of the property
37	No of Residents (Million)	Neighbourhood variable
38	Age 65 and above (% of total residents)	Neighbourhood variable
39	Resident in Condominiums and Other Apartments (% of total residents)	Neighbourhood variable
40	Condominiums and Private Flats Households (% of Total Households)	Neighbourhood variable
41	Monthly Household Income $8000 and above (% of Total Households)	Neighbourhood variable
42	Tenants (% of total households)	Neighbourhood variable
43	Uni Qualification (% of total)	Neighbourhood variable

44	CBD Km	Distance to CBD in Km
45	childcare centre km	Distance to nearest childcare centre in Km
46	community club km	Distance to nearest community club in Km
47	elderly care centre km	Distance to nearest elderly care centre in Km
48	exercise facility km	Distance to nearest exercise facility in Km
49	URA growth area km	Distance to URA growth area in Km
50	hawker centre and market km	Distance to nearest hawker centre and market in Km
51	hospital/polyclinic km	Distance to nearest hospital/polyclinic in Km
52	kindergarten km	Distance to nearest kindergarten in Km
53	library km	Distance to nearest library in Km
54	MRT station km	Distance to nearest MRT station in Km
55	park km	Distance to nearest park in Km
56	private school km	Distance to nearest private school in Km
57	Primary school km	Distance to nearest primary school in Km
58	top primary school km	Distance to nearest top primary school in Km
59	secondary school km	Distance to nearest secondary school in Km
60	junior college km	Distance to nearest junior college in Km
61	higher education school km	Distance to nearest higher education school in Km
62	shopping mall km	Distance to nearest shopping mall in Km
63	supermarket/hypermarket km	Distance to nearest supermarket/hypermarket in Km
64	bus stop km	Distance to nearest bus stop in Km
65	bus interchange km	Distance to nearest bus interchange in Km
66	No of Units	No of units in the project
67	Security	Have security facilities or not (1/0)
68	Carpark	Have Carpark facilities or not (1/0)
69	Entertainment	Have Entertainment facilities or not (1/0)
70	Chill Out	Have Chill Out facilities or not (1/0)
71	Food/Dining	Have Food/Dining facilities or not (1/0)
72	Family-Friendly	Have Family-Friendly facilities or not (1/0)
73	Fitness	Have Fitness facilities or not (1/0)
74	Function Room	Have Function Room facilities or not (1/0)

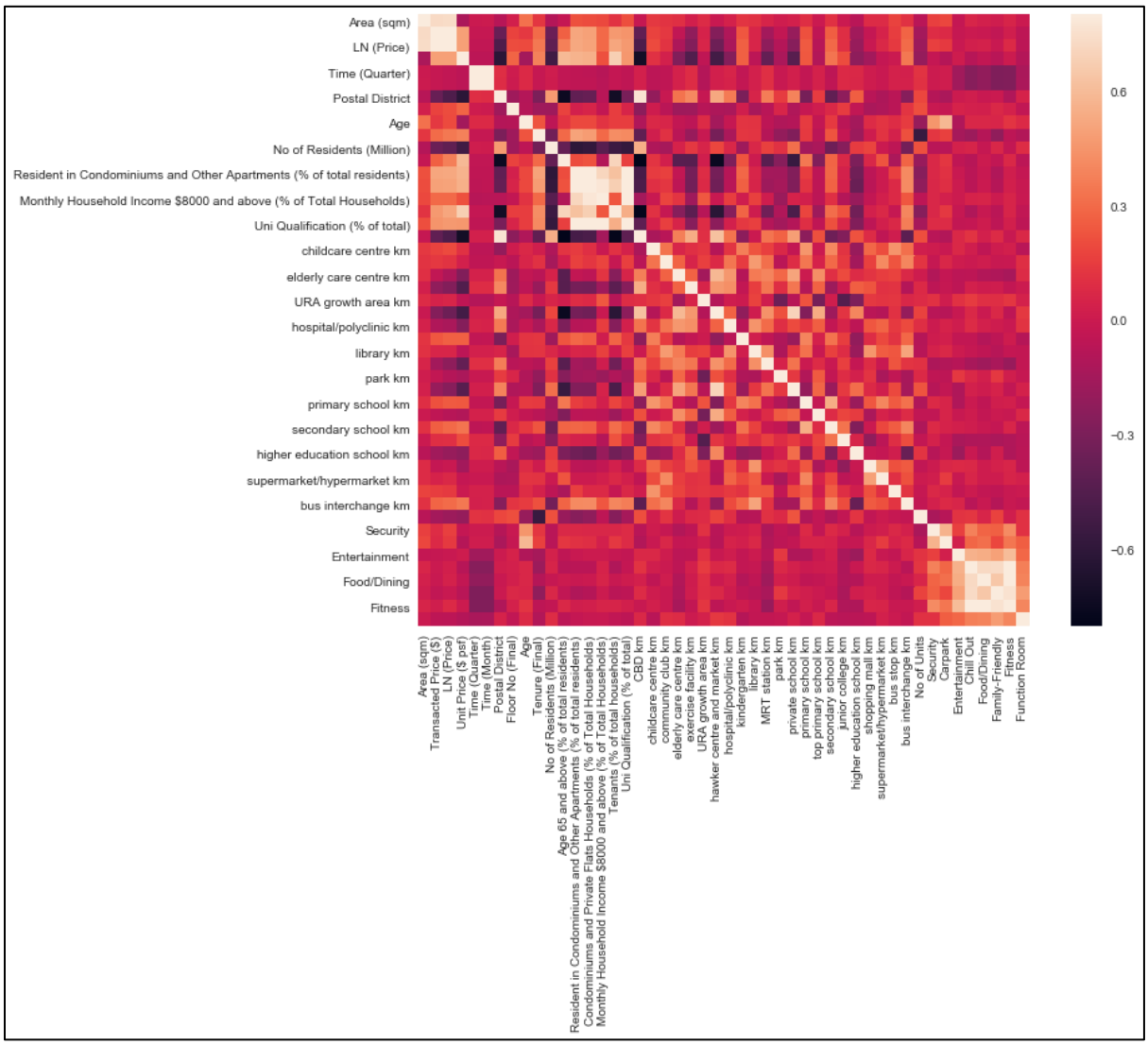
Appendix II - Basic Data Cleaning

Based on initial exploration of data, we found the need to validate data and performed following data cleaning -

1. There were several projects that had unknown completion dates, and their age was wrongly calculated as a result. To ensure the Age column has correct values, we manually found the completion dates of these 88 projects online and imputed these values into the dataset.
2. There are transactions that have more than 1 unit. In order to allow our model to be more robust and to focus on predicting property prices for transactions involving only 1 unit, we will remove transactions that have more than 1 unit.
3. Various transactions have missing values. In this case, we removed transactions with no facilities information, impute missing values for Floor No, Floor No (Final) and Unit No as zero.
4. There are several variables that would not be useful in training our models . Therefore, based on business domain knowledge we eliminated these variables - "SN", "Project Name", "Address", "No. of Units", "Type of Area", "Nett Price($)", "Unit Price ($ psm)", "Sale Date", "Address 1", "Address for Geocode", "Lease Start Date", "Latitude", "Longitude", "Completion Date", "Postal Sector", "Postal Code", "Planning Region", "Tenure", "Tenure (New)".

Appendix III - Removal of Highly Correlated Variables

We built correlation plot for all continuous variables to find highly correlated variables indicated by the intensity of the colour as shown in figure below.



Appendix IV - Top Ten Most Important Features

Apartment New Sale

	Feature Name	Importance
1	Area (sqf)	0.3577
2	Floor No (Final)	0.0955
3	childcare centre km	0.0268
4	kindergarten km	0.0231
5	bus stop km	0.0224
6	shopping mall km	0.0213
7	supermarket/hypermarket km	0.0198
8	junior college km	0.0198
9	elderly care centre km	0.0197
10	Age	0.0195

Apartment Resale

	Feature Name	Importance
1	Area (sqf)	0.1755
2	Age	0.1175
3	Floor No (Final)	0.0684
4	elderly care centre km	0.0403
5	park km	0.0365
6	CBD km	0.0317
7	bus stop km	0.0301
8	exercise facility km	0.0294
9	community club km	0.0284
10	kindergarten km	0.0276

Condominium New Sale

	Feature Name	Importance
1	Area (sqf)	0.3723
2	Floor No (Final)	0.104
3	junior college km	0.0248
4	CBD km	0.0243
5	kindergarten km	0.0237
6	top primary school km	0.0205
7	No of Units	0.0205
8	primary school km	0.0197
9	childcare centre km	0.0196
10	MRT station km	0.0189

Condominium New Sale

	Feature Name	Importance
1	Area (sqf)	0.2005
2	Age	0.0981
3	Floor No (Final)	0.0592
4	CBD km	0.0325
5	bus stop km	0.0301
6	MRT station km	0.0297
7	exercise facility km	0.0275
8	No of Units	0.027
9	secondary school km	0.0258
10	community club km	0.0252

Executive Condominium New Sale

	Feature Name	Importance
1	Area (sqf)	0.3963
2	Floor No (Final)	0.098
3	CBD km	0.0396
4	exercise facility km	0.0261
5	URA growth area km	0.0257
6	community club km	0.0251

Executive Condominium Resale

	Feature Name	Importance
1	Area (sqf)	0.1877
2	Age	0.1766
3	Floor No (Final)	0.1065
4	kindergarten km	0.0364
5	community club km	0.0312
6	CBD km	0.0299

7	park km	0.0242
8	Isunlucky Floor No	0.0229
9	childcare centre km	0.0218
10	bus stop km	0.0186

7	elderly care centre km	0.0286
8	bus stop km	0.0286
9	exercise facility km	0.0279
10	Uni Qualification (% of total)	0.0247