Analyzing housing characteristics and socioeconomic determinants of housing price in Hong Kong using tree-based regression

1. Introduction

Hong Kong is known for having one of the world's most unaffordable housing markets, characterized by soaring housing prices and increasing housing price-to-wage ratios (Leung *et al.*, 2020). Over the years, efforts by the Hong Kong government in curbing housing price escalation have included policies related to large-scale socioeconomic conditions, such as through increasing property tax or land supply (Li *et al.*, 2021). However, housing affordability is spatially heterogeneous, and there are many other specific factors that affect property prices, notably attributes of individual houses. Moreover, especially in the context of unaffordable housing, house buyers and sellers are more concerned about whether a housing price is reasonable and fair. At the same time, knowledge on factors influencing housing price can help residential developers estimate the expected return of a housing development project and conduct more reliable risk assessments (Lam *et al.*, 2008; Sawant *et al.*, 2018).

Gaining a more nuanced understanding of the determinants that affect housing prices in Hong Kong is beneficial for different stakeholders, including house buyers, sellers, residential developers, and policymakers. Therefore, this study aims to explore housing and socioeconomic factors that could potentially influence housing price in Hong Kong. This is achieved through training tree-based housing price prediction models with a diverse selection of features, then assessing the model performances as well as analyzing feature importance.

2. Literature review

Determinants of housing price

Housing prices are influenced by a diverse myriad of factors, but they can be categorized into either general or particular determinants (Shinde and Gawande, 2018). General determinants include macro-economic, social and political factors that impact overarching real estate market trends, such as the GDP and unemployment rate of a city (Lam *et al.*, 2008). On the other hand, particular determinants influence individual housing prices. In the context of

Hong Kong, examples of particular determinants include building age, distance from the Central Business District, floor of apartment (Mok *et al.*, 1995), property size (Choy *et al.*, 2007), convenient transport and distance to mass transit railway (MTR) station (So *et al.*, 1997; Yiu and Wong, 2005), and better landscape views such as garden and sea views (Hui *et al.*, 2012). Sociodemographic factors of neighbouring areas such as total population, median age and median household income also affect housing prices in Hong Kong (Li *et al.*, 2021).

Modelling studies of housing price

Previous studies undertaken on housing price prediction across different regions have employed a range of machine learning models. For instance, Song *et al.* (2022) used linear regression to model the relationship between micro-level neighborhood attributes and property price in Shanghai. Abidoye *et al.* (2019) employed support vector machines and artificial neural networks to predict Hong Kong property price index using macro-socioeconomic variables.

There have also been studies that used tree-based models to predict housing price, including decision tree and random forest regression (Gu *et al.*, 2017; Sawant *et al.*, 2018; Zhang, 2021). Decision tree regressors are advantageous over other models because they are non-parametric, therefore they can handle skewed numerical data and categorical features well (Özsoy & Şahin, 2009). Moreover, decision tree results are simple to interpret, for example, feature importance can be easily computed to analyze the relative importance of different features in influencing the target variable. Random forest is an ensemble method that aggregates the results of individual decision trees through bagging to obtain more accurate predictions (Fratello and Tagliaferri, 2019), and they work well with complicated and large datasets (Rana *et al.*, 2020).

3. Data and methods

Feature selection

Based on previous literature as well as exploration of the housing attributes listed on the Hong Kong real estate agency Midland Realty website, a list of 22 features were selected to be included in the model (Table 1). Out of the 22 variables, 17 were related to individual housing

characteristics, which contains six numerical variables and 11 were categorical. The remaining five features were socio-demographic data of the district each house is located in. For the target variable, the listing price per salable area (HKD/ft^2) was used. A large number of features were selected because it allows the exploration of a wide range of potential factors that could impact housing price in Hong Kong.

Table 1. Selected variables and their description and datatype

Variable	Description	Data type
age	Age of building [years]	Numerical
area	Salable area [ft^2]	Numerical
efficiency	Efficiency (salable area/gross floor area*100%) [%]	Numerical
floor	Floor category [low | middle | high]	Categorical
rooms	Number of rooms	Numerical
bathrooms	Number of bathrooms	Numerical
storerooms	Number of storerooms	Numerical
direction	Direction of house [N | NE | E | SE | S | SW | W | NW]	Categorical
duplex	Whether house is duplex [0 | 1]	Categorical
sea	Whether house has sea view [0 | 1]	Categorical
balcony	Whether house has a balcony [0 | 1]	Categorical
garden	Whether house has a garden [0 | 1]	Categorical
clubhouse	Whether house has a clubhouse [0 | 1]	Categorical
pool	Whether house has a pool [0 | 1]	Categorical
mtr	Whether there is an MTR station within 10 minutes walking distance from house [0 | 1]	Categorical
mall	Whether there is a mall within 10 minutes walking distance from house [0 | 1]	Categorical
park	Whether there is a park within 10 minutes walking distance from house [0 | 1]	Categorical
d_population	Total population of the district	Numerical
d_age	Median age of the district	Numerical
d_income	Median monthly domestic household income of district	Numerical
d_education	Total population aged 15 and over that attained a degree	Numerical
d_laborforce	Total labor force of the district	Numerical

Data cleaning and wrangling

The data of housing price and characteristics were obtained through web-scraping of the Midland Realty website. Apart from the 17 housing attributes, the district of each house was

also obtained. Initially, a total of 9321 samples were extracted, however, some samples contained missing data for some features. After removing all observations with missing data, the resulting dataset had 4196 observations remaining. As the scikit learn decision tree and random forest regressors (introduced below) require integer inputs for categorical features, the string label values for the features *floor* and *direction* were converted into an integer (1-3 for *floor*; 1-8 for *direction*).

Socio-demographic data by large tertiary planning units (LTPUs) was obtained from the Hong Kong 2021 Population Census. However, the LTPU boundaries differ from the housing district boundaries defined by Midland Realty. Therefore, the boundaries of each housing district were extracted from the Midland Realty website and input into QGIS alongside the LTPU boundaries and data. Area weighted average of the census data was then used to calculate the socio-demographic variable for each housing district.

Model premises

Due to the large number of categorical features in this dataset (Table 1), tree-based models are appropriate for this study because they can effectively handle categorical data, as mentioned in the previous section. Decision tree rules and feature importance can also be efficiently generated, enabling straightforward interpretation and analysis of key housing price determinants. Hence, decision tree and random forest regressors were selected for this analysis over other models.

The decision tree and random forest regressors were implemented using the *scikit-learn* library in Python, using the DecisionTreeRegressor and RandomForestRegressor respectively. Both models have a list of hyper-parameters that govern model behavior and performance. For decision tree, key hyper-parameters include the maximum depth of the decision tree (max_depth) and the minimum number of samples required to split a decision node (min_samples_split) and to be at a leaf node (min_samples_leaf). Random forest contains two additional important hyper-parameters: the number of trees (n_estimators) and the maximum number of features that are considered during each split (max_features). Limiting the number of features sampled considered ensures that the individual decision trees are not correlated, which improves model performance by reducing overfitting. Hyper-parameter

tuning was conducted using k-fold cross validation. Five folds were used for tuning both models. The resulting hyper-parameter values used in this study are presented in Table 2.

Table 2. Hyper-parameter values used for the decision tree and random forest models used in this study (discussed below).

Hyper-parameter	Decision tree	Decision tree (log)	Random forest	Random forest (log)
max_depth	12	10	12	12
min_leaf_samples	1	1	1	1
min_leaf_split	17	7	7	7
n_estimators	N/A	N/A	300	200
max_features	N/A	N/A	0.5	0.5

4. Results and discussion

Exploratory data analysis

Before fitting the models, exploratory data analysis was undertaken to understand the relationships between the features and the target variable, which could help interpret the model results. Figure 1 shows scatter plots of housing price against eight selected numerical features. For income and d_area, the plots show a positive correlation between these features and the price, where an increase in saleable area and median household income of the district results in higher housing price. Smaller building age (age), as well as smaller total district population (d_population), population that attained a degree (d_education) and labor force (d_laborforce) generally corresponds to higher housing price. On the other hand, efficiency and district median age do not seem to exhibit a clear relationship with housing price.

Figure 2 presents the distribution of the observation values for housing price and all variables. It is worth noting that housing price and many of the numerical variables are right skewed (Figure 3a). Even though decision tree models are non-parametric and generally not impacted by uneven distribution of variables, transforming the data to reduce skew could sometimes improve model performance. Hence, log transformation was applied to those variables (Figure 3b) to run the models as well to test if reducing skew would improve model performance. Furthermore, the features duplex and garden were heavily imbalanced with

almost all observations falling into “0” (not duplex, no garden). This could impact the model ability to account for the influence of these features on housing price.

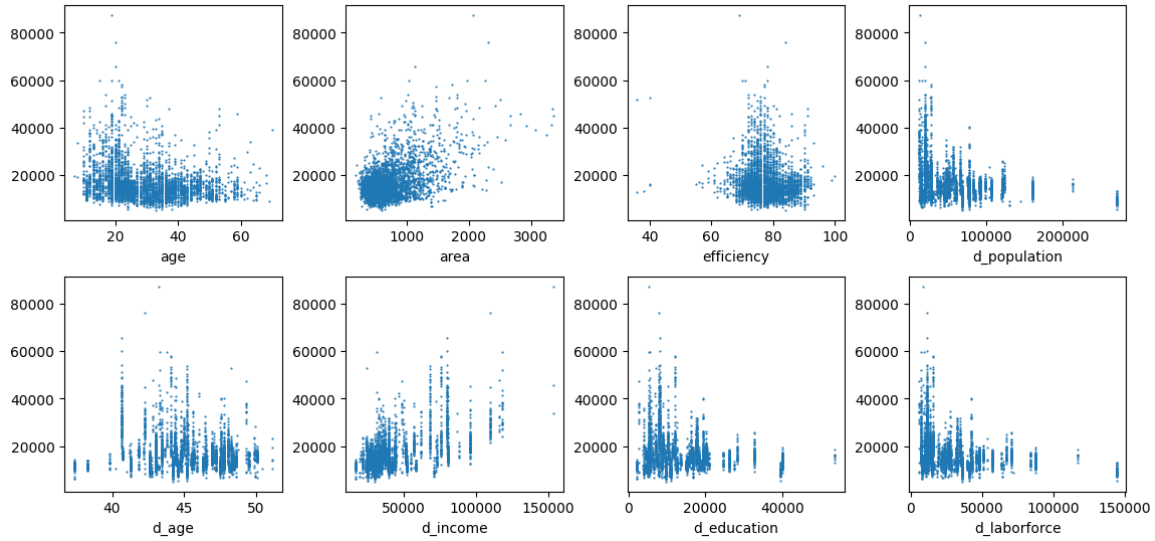

Figure 1. Scatter plots showing the relationships between selected numerical features and housing price.

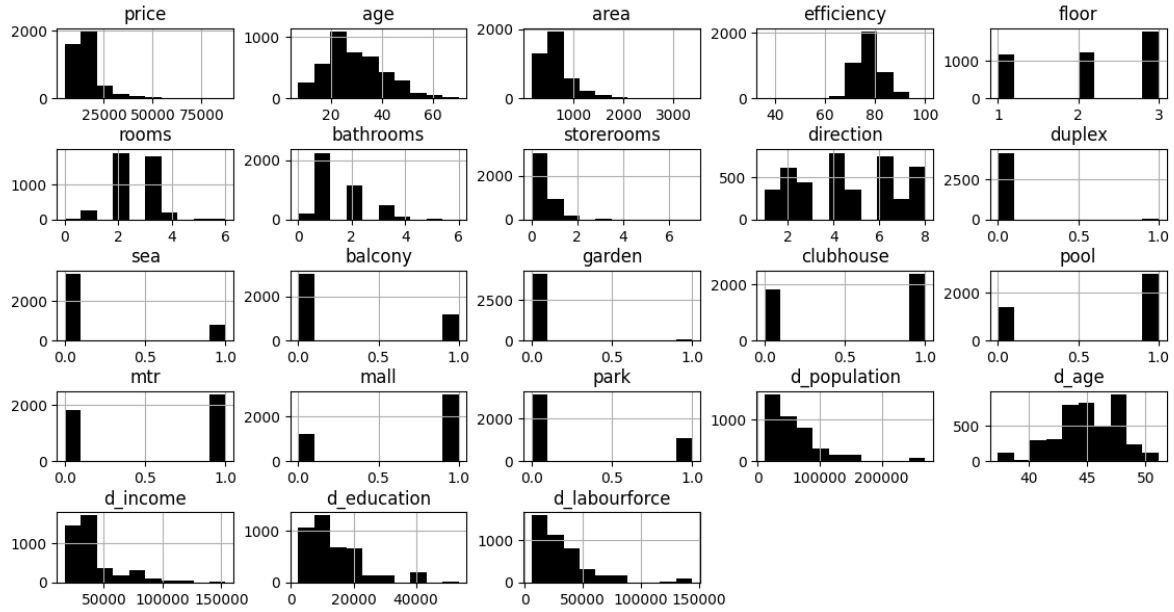

Figure 2. Histograms showing the distribution of housing price and all features.

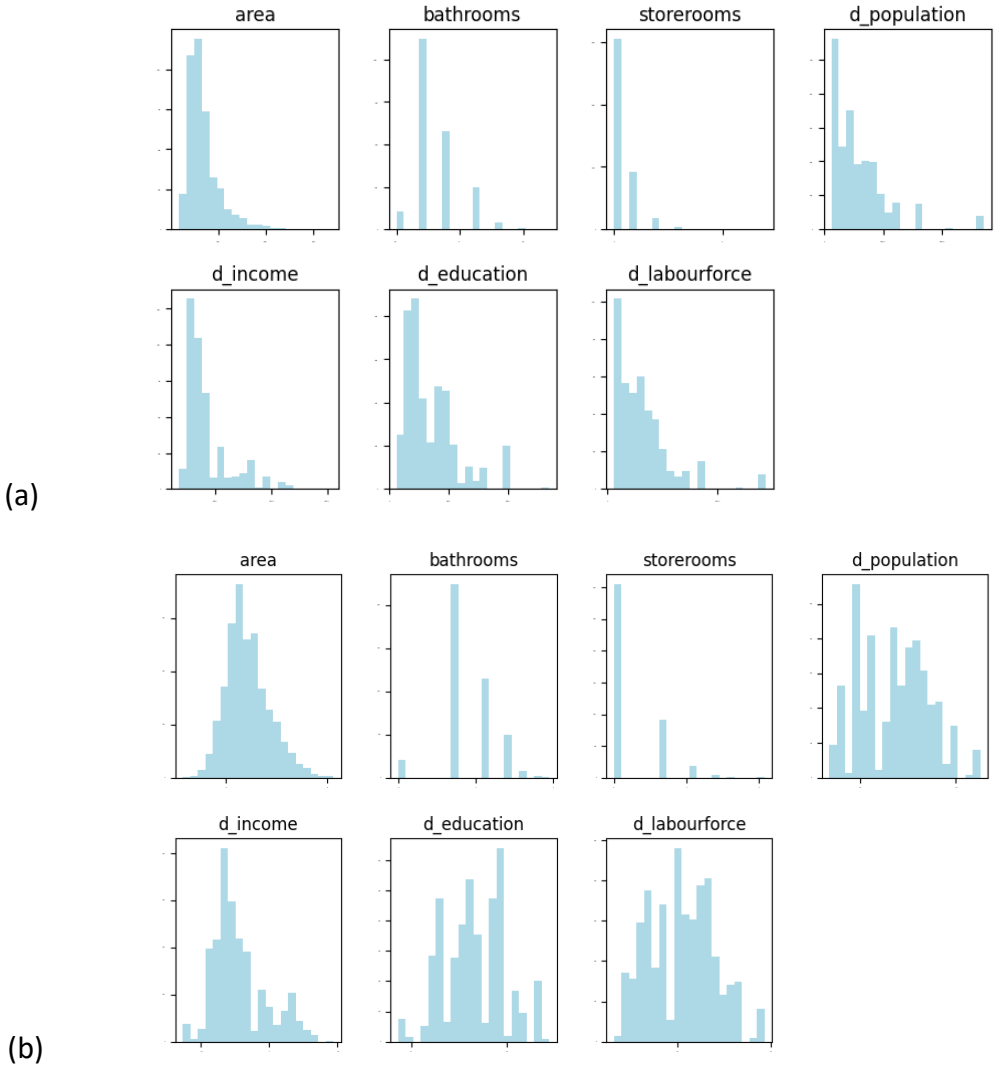

Figure 3. (a) Features that are right skewed and (b) the distribution of the features after applying log transformation.

Comparison of model performance and results

To assess each model's performance, R2 and root mean squared error (RMSE) were used as evaluation statistics. R2 measures the goodness of fit of a model, which shows how well a model can predict unseen samples. An R2 of 1 represents a perfect fit of the model, therefore an R2 closer to 1 reflects generally more robust model performance. RMSE quantifies the average error between the predicted and actual values, so a smaller RMSE indicates better model fit and performance. Table 3 presents the R2 and RMSE for the four models used in this study: the baseline decision tree, decision tree using log transformed data, random forest and random forest using log transformed data.

Table 3. Model performance of the tree-based models used in this study.

Model	R2	RMSE
Decision tree	0.7576	3608.3914
Decision tree (log)	0.7725	3495.5654
Random forest	0.8298	3023.6948
Random forest (log)	0.8290	3030.8332

The random forest models resulted in better model performance compared to the decision tree models. Comparing the baseline decision tree and random forest models, the R2 increased from 0.7576 to 0.8298, showing an improved ability of the model to predict unseen data. RMSE decreased by 579.6966, which shows that the mean error between all predictions and actual values decreased by around 580 HKD/ft^2. Similarly, the two models ran using the log transformed data also demonstrated better model performance for random forest, with higher R2 and lower RMSE. Furthermore, the effect of reducing the skew of the data was not significant in improving model performance. For decision tree, using log transformed data only resulted in a slight improvement in R2 of 0.0149, while R2 dropped by 0.0008 when using the transformed data for random forest. Therefore, it can be concluded that the uneven distribution of the dataset did not have a significant impact on model accuracy.

Overall, the baseline random forest model performed the best with the highest R2 (0.8298) and lowest RMSE (3023.6948). This R2 value is comparable to similar housing price prediction studies using different models, such as Lam *et al.* (2008) which employed artificial neural networks to predict property price in Hong Kong using a list of particular housing price determinants similar to this model as well as several macro-economic determinants, and resulted in R2 values of around 0.8. This further shows that the random forest model in this study is reliable and can be used to analyze feature importance in the following section.

Feature importance

From Figure 4, out of all the variables, housing price in Hong Kong is mostly influenced by the median household income of the district where the house is located in. From Figure 1, it can be seen that wealthier neighborhoods result in higher housing prices. Other important determinants of housing price include the salable area and building age, in which larger salable areas and newer buildings are more expensive (Figure 1). On the contrary, most

categorical features, such as duplex, park, mall and garden, have negligible effects on housing price in Hong Kong, even though this could be influenced by the imbalance in feature data (Figure 1). Out of all categorical variables, the accessibility of a nearby MTR station and having a sea view are the most important in influencing housing price, which corroborates with previous findings. However, since all variables related to the district are among the highest ranking in feature importance, it can be implied that the geographical location of the house is the most important factor that determines the price of a house in Hong Kong.

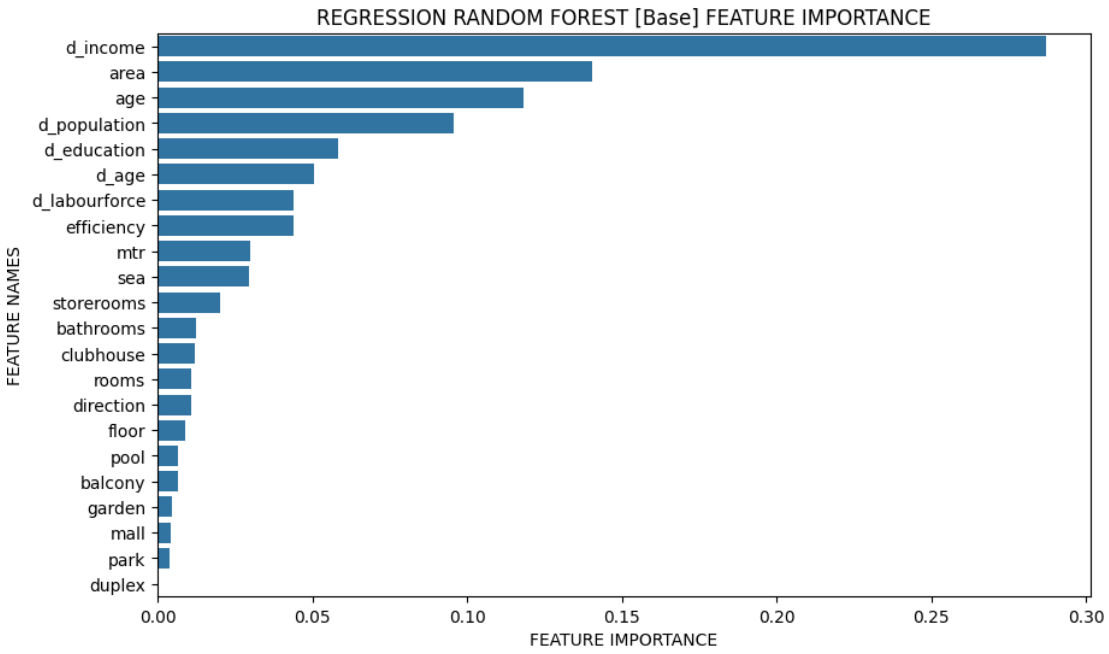

Figure 4. Bar graph showing feature importance based on the random forest model.

5. Conclusion

Amongst the tree-based regression models compared in this study, random forest had the best model performance with the highest R2 and lowest RMSE. Feature importance analysis showed that geographical location is the most important determinant of housing price, followed by the salable area and building age of the house. Future modelling research can be done using historical housing transaction data and incorporating general macro-level determinants as well (Sawant *et al.*, 2018), which enables future projections of housing price in Hong Kong.

References

Abidoye, R.B., A.P.C. Chan., F.A. Abidoye. and O.S. Oshodi. 2019. Predicting property price index using Artificial Intelligence Techniques. International Journal of Housing Markets and Analysis 12(6): 1072–1092.

Choy, L.H., S.W. Mak. and W.K. Ho. 2007. Modeling Hong Kong real estate prices. Journal of Housing and the Built Environment 22(4): 359–368.

Fratello, M. and R. Tagliaferri. 2019. Decision Trees and Random Forests. In S. Ranganathan, M. Gribskov, K. Nakai and C. Schönbach (Eds) Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, Amsterdam: Elsevier.

Gu, G. and B. Xu. 2017. Housing market hedonic price study based on boosting regression tree. Journal of Advanced Computational Intelligence and Intelligent Informatics 21(6): 1040–1047.

Hui, E.C.M., J.W. Zhong. and K.H. Yu. 2012. The impact of landscape views and storey levels on property prices. Landscape and Urban Planning 105(1–2): 86–93.

Lam, K.C., C.Y. Yu. and K.Y. Lam. 2008. An artificial neural network and entropy model for Residential Property Price Forecasting in Hong Kong. Journal of Property Research 25(4): 321–342.

Leung, C.K., J.C. Ng. and E.C. Tang. 2020. Why is the Hong Kong housing market unaffordable? some stylized facts and estimations. Federal Reserve Bank of Dallas, Globalization Institute Working Papers April 2020.

Li, J., W. Fang., Y. Shi. and C. Ren. 2021. Assessing economic, social and environmental impacts on housing prices in Hong Kong: A time-series study of 2006, 2011 and 2016. Journal of Housing and the Built Environment 37(3): 1433–1457.

Mok, H.M., P.P. Chan. and Y.-S. Cho. 1995. A hedonic price model for private properties in Hong Kong. The Journal of Real Estate Finance and Economics 10(1): 37–48.

Rana, V.S., J. Mondal., A. Sharma. and I. Kashyap. 2020. House price prediction using optimal regression techniques. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) 18 December 2020, 203–208.

Sawant, R., Y. Jangid., T. Tiwari., S. Jain. and A. Gupta. 2018. Comprehensive analysis of housing price prediction in Pune using multi-featured random forest approach. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) August 2018.

Shinde, N. and K. Gawande. 2018. Survey on predicting property price. 2018 International Conference on Automation and Computational Engineering (ICACE) October 2018, 1–7.

So, H.M., R.Y.C. Tse. and S. Ganesan. 1997. Estimating the influence of transport on House prices: Evidence from Hong Kong. Journal of Property Valuation and Investment 15(1): 40–47.

Song, Q., Y. Liu., W. Qiu., R. Liu. and M. Li. 2022. Investigating the impact of perceived micro-level neighborhood characteristics on housing prices in Shanghai. Land 11(11): 2002.

Yiu, C.Y. and S.K. Wong. 2005. The effects of expected transport improvements on housing prices. Urban Studies 42(1): 113–125.

Özsoy, O. and H. Şahin. 2009. Housing price determinants in Istanbul, Turkey. International Journal of Housing Markets and Analysis 2(2): 167–178.

Zhang, Z. 2021. Decision trees for objective house price prediction. 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) December 2021, 280–283.