House Price Prediction using Random Forest

Anojan Satheesnathan*, Sankeerthan Kasilingam† and Sampath Deegalla‡
Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka
Email: *anojan010@gmail.com, †shankeerthan1995@gmail.com, ‡dsdeegalla@pdn.ac.lk

***Abstract*—Housing price is influenced by multiple factors such as location, house size, number of bathrooms and number of bed rooms.Identifying all factors is crucial. This research can help house seller or a real estate agent to make better-informed decisions in the housing market. There is no work reported on the use of machine learning algorithms to predict the house prices in Sri Lanka. This paper presents research predicting house prices with a data set collected from online house-selling platform. This study focused on Colombo district the capital of Sri Lanka.According to our results, after removing outliers by clustering, the Random Forest model performance was improved. Along with prediction this paper analyses other similar researches done previously, it explores clustering to remove outliers, identifies feature importance and compares and discusses the results.**

***Index Terms*—house value prediction, machine learning, regression , random forest**

I. Introduction

The real estate has been attractive and competitive market for long years because price decline of real estate properties are very rare compared to other assets [1] [2]. An accurate assessment of houses and real estate properties are very important to many parties such as local government, house owners, developers, investors, appraisers, mortgage lenders and insurers [3]. Before the arise of Machine Learning the house prices often estimated by professional appraisers but it had few drawbacks. The traditional services are expensive and appraiser is likely to vest interest to one party. When it comes to accuracy of estimation, the machine learning models performs better than hedonic models [4].

Similar studies using Machine Learning to predict house prices were done in different countries, but were totally different in terms of algorithms, features, parameters and data set [5] [6] [4] [7] [8] [9].

In Machine Learning context, the house price is characterized as a set of attributes. The common attributes that affect house prices are house size, location and number of bedrooms. But these are not limited to physical factors [2] [5] there are so many external factors affect the house prices such as economy, GDP growth and etc [5].

It's crucial to build a good Machine Learning model to predict house price. The data has to be studied carefully by feature extraction, prep-processing and parameter tuning [10].

The paper is divided into the following sections. Section two provides an overview of background research, section three introduces the data set used, section four presents methodology and section five provides the results of the experiments and discussion on the results.

II. Related Work

Our study inspired from [5], the authors collected data from an online property platform called idealistica.com. The data set consisted of 2,266 instances with 20 factors. The data set is limited to Salmanca district, Madrid, Spain. They implemented Support Vector Regression, K-Nearest Neighbours, Ensemble of Regression Trees and Multi Linear Perceptron models. According to their findings the model built with Ensemble of Regression trees performs better. Their model scored 16.80% mean absolute error against mean prices and 5.71% median absolute error against median prices.

Another interesting study was done with online housing advertisements in Brazil. The data set consisted of 12,223,58 housing advertisements with 24 features and this was collected from 2015 to 2018. The study compares two ML approaches; the Random Forest based model and the Deep Learning model. According to their findings both models perform well.However, the Random Forest based model performs better with numeric attributes [11].

A study was carried out in Malaysia with real house data set of 2016. They implemented the model with Random Forest, Decision Tree, Ridge, Linear and Lasso Regression. They evaluated and compared based on R2 score and RMSE. According to their findings the Random Forest model performs better than others [10].

Another study was done in South Korea that compares Random Forest based model with traditional ordinary Least Square method. The author used a real data set of 16,601 samples within the period of 2006 to 2017. The study took account of structural, neighbourhood, locational and macro attributes such as transaction period, GDP and mortgage interest rate. They used 1:9 split of data set for test and train and considered averages of 10 experiments to reduce the possibility that results occurred by chance. They found only 5.5% average deviation from actual price as compared to 20% using the OLS model. They noted the Random Forest model can perform better than traditional models [12].

A study was done on single family open house data set of Arlington Virginia, USA. The data set comprises of 27,649 houses and factors such as lot size, years built, zip code and location. The authors bench marked Linear Regression and compared it to the Random Forest model. From their results Random forest performs better than Linear Regression [13].

From mentioned studies, we could see that the studies used

either real data set or data set collected from advertisements. The studies used Random Forest model and got reasonable performance, and they were done in different countries. It's useful to do research on the local market (Sri Lanka) and to see how the machine learning models can be utilized for house price prediction.

III. Data

"Tab. I" shows the description of the collected data set. Location is the only categorical attribute which contains 40 different locations and others are numerical attributes.

For our study we only took into account the market in the Colombo district, the capital of Sri Lanka. To establish data pool for our study, we collected the data within the period of July 2019 to April 2020 from the website ikman.lk which is a leading online selling and buying platform in Sri Lanka. It allows sellers to post house advertisements on the website. For our study we assumed the advertised house prices as actual sold prices. The data set consisted of 12,712 houses with five attributes which are the number of bedrooms, number of bathrooms, house size, land size, and location.

IV. Methodology

A. Algorithms

K-Means: It is simplest unsupervised algorithm and it identifies k (predefined) number of clusters (non-overlapping instances) based on distance between instances [14].

Random Forest Regression: It is a supervised algorithm and it uses ensemble method bagging. It builds independent decision trees for sub sample data and make predictions by combining independent models [15].

B. Data Cleaning and Preprocessing

We found that 5,164 number of duplicates exists in collected data set and we removed them. Along with removing duplicates we dropped unstable data instances which has null value for any fields. We established a consistent stable data set of 7,547 instances. We used one-hot-encoding to convert categorical locations. To make the model learn more about data, we derived three new attributes from existing attributes [16]. They are the ratio between house size and land size, ratio between number of beds and number of baths and bed size. They bring the hidden information and relationships between existing attributes. The ratio between land size and house size was used to separate single story houses. Ratio below 1.0 were considered as single story houses. The bed size was derived with an assumption of the total bedrooms' size take the whole space of a house.

C. Data Exploration and Analysis

In data related studies, the visualization of data is important to understand complex problems [17]. We visualized the data using pair plot to analyse the relationships between variables. From the pair plot we found out that there is no linear relationship between house price and other independent variables. Only the number of beds and number of baths shows an approximate linear relationships.

TABLE I
Data Description

Attribute	Type	Mean	Std
Price (LKR)	int64	3.03e+07	3.99e+07
Land size (Perches)	float64	34.11	1843.47
Number of beds	int64	3.93	1.25
Number of baths	int64	2.99	1.28
House size	float64	2693.61	2663.46
Location	String	NA (eg: Piliyandala)	NA

The location is a key attribute in determining house prices [18]. The collected data set covers 40 regional zones in Colombo. Out of 40 zones, six zones have lesser than 20 instances and the maximum number of instances were recorded for zone 'Piliyandala' which is 1448.

D. Outliers Detection

The outliers are the houses that do not follow the crowd (majority instances). These instances show deviation from majority of instances in terms of values of attributes. These instances may be outliers by natural or outliers due to incorrect attribute values. To fit a model to data (majority instances), the outliers should be removed. Otherwise they will introduce disturbances to model [19]. In our study, the assumption (advertised prices are considered as sold prices) created more outliers and introduced difficulty to remove outliers with typical process.

We found that some of the houses have extreme values for some attributes. To filter them, we analysed and derived threshold values (upper and lower). Some houses have large land size, and the price is not only for the houses and it includes the land as well. We found the land size below 30.0 perches contain 97.53% instances. Tab. II shows the threshold values with respective attributes.

The clustering is the technique to group similar instances. By considering clustering to remove outliers, have been studied, and a study propose clustering based on key attributes is a method to remove outliers and the author summarized K-Means clustering performed better than distance based outlier algorithms [20]. We considered the K-Means algorithm to cluster the data in order to eliminate and identify similar groups of house instances. We used elbow method to decide the optimal number of clusters. We found three number of clusters correctly identify the outliers (the two small clusters of house instances). The results are presented in the result and discussion section.

TABLE II
Threshold value of Attributes

Attribute	Lower Threshold	Upper Threshold
Price(LKR)	8,000,000	50,000,000
House size (sqft)	0	20000
Land size(Perches)	0	30

E. Model Setup

We used sckitlearn library to implement models. For each experiment, we splitted the data set into 1:3 ratio for test and train. For the initial experiments with any model, we used default parameters provided by sckitlearn except number of trees (n-estimators). We built multiple Random Forest models in range of estimators. With whole data set, we tested all possible combinations of feature sets and identified most importance feature sets. Then every model we tested with the importance feature sets and evaluated based on the evaluation metrics.

F. Evaluation Metrics

We considered RMSE (Root Mean Squared Error) and R2 score (Coefficient of Determination) to evaluate and compare models. The range of R2 score is 0.0 to 1.0 but it can take negative values in worst case.

$$\text{RMSE} = \frac{\sum_1^n (y\_pred - y\_test)^2}{n} \quad \text{R2} = 1 - \frac{\sum_1^n (y\_test - y\_pred)^2}{\sum_1^n (y\_test - mean(y))}$$

y_test: test target, y_pred: predicted target, n: number of test instances, y: all target values

G. Parameter Tuning

We used gridSearch (sckit implementation) to tune hyper parameters of Random Forest regressor. After we gained a certain level of confidence in the model by changing the parameters manually, we used grid search to tune in order to find optimal parameters.

V. Result and Discussion

In this section, the results of experiments and discussions are presented. The experiments include both outlier detection model and prediction model.

The Initial experiment was about building a Linear Regression model to identify relationships between variables. We conducted the experiment with whole dataset and single story houses. We got positive R2 score for both models but the RMSE values were very high (24 Million and 19 Million). And we noticed our Linear Regression model scored high RMSE for whole houses' data set compared to single story houses . We found one possible reason could be the single story houses have similar features. We concluded that the independent variables do not have a linear relationship with house price. The plot with each and every variables also confirmed non-existence of linear relationship.

Like the Linear Regression model, we conducted the experiments with both whole data set and single story houses using Random Forest regression algorithm. "Tab. III" shows the results. Similar to Linear Regression model, the single story house model scored lower RMSE compared to whole data set model. The mean prices of single story houses and whole houses were 24 Million and 28 Million respectively. The same reason that the similar features between single story houses could affect single story model performance. We noticed the RMSEs were very close to the mean prices. The assumptions that we made (sold price = advertised price) and existence of outliers affect the models worsely.

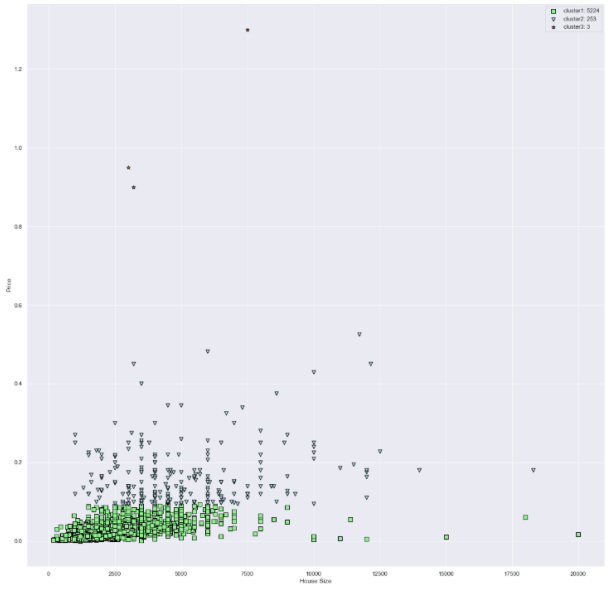

Fig. 1. The Price against House Size, Cluster of 3

To further experiment we found out a list of important feature sets from all possible combinations. We used the Random Forest Regression model to evaluate importance in terms of model scores R2 and RMSE. The idea was to remove or avoid unnecessary trials in hyper parameter tuning.

To eliminate outliers further, we implemented K-Means model with whole dataset. We conducted the experiments in two ways by changing number of clusters and features. To find out optimal number of clusters, we used elbow method. From the curve of SSE (Sum of Squared Error) against number of clusters, we noticed after two clusters SSE declines slowly. From our data set context, splitting further (more than three clusters) would reduce number of instances per clusters. Therefore we decided and did the experiments with up to 5 clusters.

"Tab. IV" shows the clustering with and without prices, number of clusters and respective results of Random Forest regression models. It only shows the clusters with large number of instances. From the results, we concluded that considering the variable House Size, Beds, Price and the variables in clustering would result in a better clustering model. In terms of metrics (R2, RMSE), the model with clustered data (removed outliers with clustering) performed better than previously built models.

TABLE III
RESULTS OF RANDOM FOREST MODEL BEFORE CLUSTERING

Data set	R2		RMSE (Million)	
	Min	Max	Min	Max
Whole houses	0.04	0.40	24	30
Single story houses	0.19	0.36	21	24

TABLE IV
RESULTS OF RANDOM FOREST MODEL AFTER CLUSTERING

Clustering	Number of Clusters	R2	RMSE(Million)
Not Considered price	2	0.28	10
	3	0.20	9.7
Considered price	2	0.68	8.9
	3	0.70	5.2

VI. CONCLUSION

The house price prediction using Machine Learning shows a thriving trend. In this paper we explored the way to predict house prices using Random Forest regression with an approximate data set.

We collected data from an online advertising platform and we made an assumption that the advertised prices are the same as sold prices. We cleaned data, removed outliers and extreme values' instances. We used K-Means algorithm to detect outliers. We concluded that, the clustering techniques works well in detecting outliers, where an instance has high degree of chance to be an outlier. We found out the assumption we made and undetected outliers causes Radom Forest model performance. Out of there, the best models scored RMSE(average of 10 trials) of LKR 6 Million.

Our study was done with some limitations, such as approximate data set and data set is limited to a small period of time. We only considered the Random Forest regression to build models. To further explore, it is useful to consider real data set, time series analysis and deep learning models. Considering a wide range of factors (GDP, GDP growth rate, per capita, etc.) along with time series data would improve performance of the models.

REFERENCES

[1] R. Manjula, S. Jain, S. Srivastava, and P. Rajiv Kher, "Real estate value prediction using multivariate regression models," in *Materials Science and Engineering Conference Series*, vol. 263, p. 042098, 2017.

[2] N. Shinde and K. Gawande, "Survey on predicting property price," in *2018 International Conference on Automation and Computational Engineering (ICACE)*, pp. 1–7, IEEE, 2018.

[3] H. Xu and A. Gade, "Smart real estate assessments using structured deep neural networks," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–7, IEEE, 2017.

[4] V. Limsombunchai, "House price prediction: hedonic price model vs. artificial neural network," in *New Zealand agricultural and resource economics society conference*, pp. 25–26, 2004.

[5] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *Applied Sciences*, vol. 8, no. 11, p. 2321, 2018.

[6] A. Ng and M. Deisenroth, "Machine learning for a london housing price prediction mobile application," *Imperial College London*, 2015.

[7] J. Wang, S. Hu, X. Zhan, Q. Luo, Q. Yu, Z. Liu, T. P. Chen, Y. Yin, S. Hosaka, and Y. Liu, "Predicting house price with a memristor-based artificial neural network," *IEEE Access*, vol. 6, pp. 16523–16528, 2018.

[8] O. Kitapci, Ö. Tosun, M. F. Tuna, and T. Turk, "The use of artifi cial neural networks (ann) in forecasting housing prices in ankara, turkey," *Journal of Marketing and Consumer Behaviour in Emerging Markets*, no. 1 (5), pp. 4–14, 2017.

[9] V. Chiarazzo, L. Caggiani, M. Marinelli, and M. Ottomanelli, "A neural network based model for real estate price estimation considering environmental quality of property location," *Transportation Research Procedia*, vol. 3, pp. 810–817, 2014.

[10] Y. F. Chang, W. C. Choong, S. Y. Looi, W. Y. Pan, and H. L. Goh, "Analysis of housing prices in petaling district, malaysia using functional relationship model," *International Journal of Housing Markets and Analysis*, 2019.

[11] B. Afonso, L. Melo, W. Oliveira, S. Sousa, and L. Berton, "Housing prices prediction with a deep learning and random forest ensemble," in *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pp. 389–400, SBC, 2020.

[12] J. Hong, H. Choi, and W.-s. Kim, "A house price valuation based on the random forest approach: the mass appraisal of residential property in south korea," *International Journal of Strategic Property Management*, pp. 1–13, 2020.

[13] C. Wang and H. Wu, "A new machine learning approach to house price estimation," *New Trends in Mathematical Sciences*, vol. 6, no. 4, pp. 165–171, 2018.

[14] Y. Li and H. Wu, "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.

[15] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. Apr, pp. 1063–1095, 2012.

[16] M. Cocea and S. Weibelzahl, "Log file analysis for disengagement detection in e-learning environments," *User Modeling and User-Adapted Interaction*, vol. 19, no. 4, pp. 341–385, 2009.

[17] M. N. Sadiku, A. E. Shadare, S. M. Musa, and C. M. Akujuobi, "Data visualization," *International Journal of Engineering Research And Advanced Technology (IJERAT)*, vol. 2, no. 12, pp. 11–16, 2016.

[18] L. Fernández-Durán, A. Llorca, N. Ruiz, S. Valero, and V. Botti, "The impact of location on housing prices: applying the artificial neural network model as an analytical tool.," 2011.

[19] T. Wang and Z. Li, "Outlier detection in high-dimensional regression model," *Communications in Statistics-Theory and Methods*, vol. 46, no. 14, pp. 6947–6958, 2017.

[20] A. Christy, G. M. Gandhi, and S. Vaithyasubramanian, "Cluster based outlier detection algorithm for healthcare data," *Procedia Computer Science*, vol. 50, pp. 209–215, 2015.