Stacking Ensemble Learning for Housing Price Prediction: a Case Study in Thailand

Conference Paper · January 2021

DOI: 10.1109/KST51265.2021.9415771

5 authors, including:

Ekapol Chuangsuwanich
Massachusetts Institute of Technology
61 PUBLICATIONS **725** CITATIONS

SEE PROFILE

Stacking Ensemble Learning for Housing Price Prediction: a Case Study in Thailand

Gan Srirutchataboon
Home Dot Tech. Co., Ltd.
Bangkok, Thailand
gan@homedottech.com

Saranpat Prasertthum
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
saranpat.pr@student.chula.ac.th

Ekapol Chuangsuwanich
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
ekapol.c@chula.ac.th

Ploy N. Pratanwanich*
Department of Mathematics and Computer Science
Chulalongkorn University
Bangkok, Thailand
naruemon.p@chula.ac.th

Chotirat Ratanamahatana
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
ann@cp.eng.chula.ac.th

***Abstract*—In this paper, we analyze the housing price data obtained from a leading Thai real estate website and Open Street Maps (OSM) to identify the features that affect the housing price in Thailand from 2015 to 2019. Moreover, we propose a model based on a stacking ensemble learning framework, where the predictions are generated by stacking three base learning models consisting of a convolutional neural network (CNN), an ensemble model (such as random forests (RF), extreme gradient boosting (XGBoost) and adaptive boosting (AdaBoost)) and a simple linear regression technique. The CNN is used to extract features from house images which are then combined with traditional features to estimate the initial price. The prediction is then calibrated using linear regression. Compared to individual models, the proposed model achieves a Mean Absolute Percentage Error (MAPE) of 17.83%, significantly outperforming other baselines.**

***Index Terms*—stacking model, CNN, housing price prediction, machine learning, linear regression**

I. Introduction

Machine learning is now being utilized to improve business capabilities and finding insights through data analysis for companies across the globe. For real estate, the development of a housing price prediction model can assist developers and customers to make proper decisions when making a purchase. Moreover, an accurate price prediction can estimate future economic growth in each country. Therefore, developing an effective model for predicting housing price has been becoming a crucial issue and receiving much attention [1], [2].

So far, there are many works focusing on house price prediction. For instance, D. Banerjee *et al.* [4] applied several ML techniques to predict the direction of the price, which the performances are measured by four parameters including accuracy, precision, specificity and sensitivity. In [5], S. Lu *et al.* proposed a hybrid regression technique to predict the price of an individual house. Their final model is a weighted combination of various ML models such as least absolute shrinkage and selection operator (Lasso) [6], Ridge regressions and XGBoost models. Furthermore, the historical property data transactions in Australia were analysed to discover useful predictive models in [7]. They showed that the combination between stepwise technique and support vector machine (SVM) [8] model could provide an acceptable performance. A. Varma *et al.* [9] utilized various regression techniques to evaluate the real-time housing price for various localities around Mumbai, India.

In this paper, we focus mainly on the dataset of 2^{nd} hand single houses in Thailand, a region which is receiving much attention from many local and foreign investors. The dataset was obtained from http://www.home.co.th, a leading Thai real estate website, which includes structured information such as the size and number of rooms and unstructured information such as images of the property. Moreover, we also utilize information from Open Street Map (OSM) to identify places that might affect the price of the surrounding areas. To extract useful information from the dataset as well as performing features selection, we first provide insights obtained from our data analysis and visualizations. According to our analysis, the housing price in our dataset depends on several intrinsic attributes such as the number of bedrooms and sizes. Furthermore, external attributes such as geographical location and the number of the roads around the house can also be used to predict its price.

Recently, a stacking model framework is widely used to improve the accuracy of any single model even further by cascading different ML models in a series. However, the problem is how to select models for the framework that are appropriate to the data. Here, CNN network was used as a trainable feature extractor to provide the input for an XGBoost

*Corresponding author

978-1-7281-7602-4/21/$31.00 ©2021 IEEE

model. This technique outperformed other techniques on the same dataset.

In our study, we aim to evaluate the effectiveness of stacking different kinds of models, namely CNN-random forests, CNN-XGBoost and CNN-AdaBoost for finding the most effective stacking model for our dataset. Moreover, we found that a simple linear regression model can be used to calibrate the final prediction which can reduce the gap between the actual and the previously predicted values even further.

The rest of this paper is organized as follows. In Section II, we describe the dataset and perform several kinds of data analysis. The concept of conventional stacking models are described in section III(a). We descrbie the proposed stacking model with linear regression technique in section III(b). Our experimental results are demonstrated in section IV, and finally we conclude in section V.

II. Data Explanation and Exploration

A. Dataset

In this paper, we utilize the dataset of 2^{nd} hand single houses in Thailand, which was advertised on www.home.co.th, a leading real estate website in Thailand. As shown in Table I, each attribute in the dataset is described, which consists of 'YearBuilt', 'LotSize','Baths', 'Beds', 'Parking', 'DistrictID', 'Size', 'ListingID', 'SalePrice', 'Lat', 'Lng', 'AvgMinSalePrice', and 'AvgDistrict'. The distinction among the advertisements house is made by attribute 'ListingID'. The information from OSM, which related to the location of the house including the number of convenience store such as seven-eleven (7-11), Max-Value (MiniShop), the number of coffee shop as Starbucks (Starbucks), the number of the road (Roads) and the number of public transport station (PubTran). Those are counted within 1 kilometre (km.) from the house.

Fig. 1. Example images from the dataset

As we mentioned earlier, we utilize the image set for generating features using CNN [3]. An input image will be normalized by dividing by 255. Therefore, the data is scaled at an interval of 0 to 1. The example of the images houses in our dataset are shown in Fig 1.

B. Exploratory Data Analysis

This section, we explore the dataset using multiple exploratory techniques to find uncover underlying patterns and select features. In Fig. 2, Spearman correlation matrix [11] between pairs attribute in Table I is presented. The coefficient ranges between -1 and 1, where 1 means two attributes have

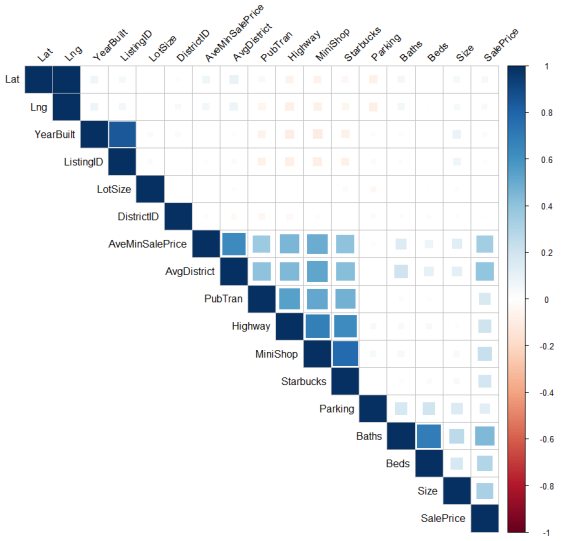

Fig. 2. correlation matrix between attributes.

strong and positive association. In contrast, -1 means two attributes have strong and negative association. As shown in Fig. 2, attribute 'SalePrice' has strong positive relationship with attributes 'Beds', 'Baths', 'LotSize', 'Size', 'AvgDistrict' and 'AvgMinSalePrice'.

Further, Fig.3 presents the probability density function (PDF) of attribute 'Beds' over three price ranges. Fig. 3(a) highlights that the houses with price lower than 3 million Thai Baht (THB) tend to have 2 or 3 bedrooms. However, there are proportion of 4 bedrooms houses in this price range, but their price are around 2.5 million THB. If consider the price ranges between 3 to 12 million THB, the proportion of 3 bedrooms houses is dominant. While there is a small proportion of 4 bedrooms houses from the price between 3 to 10 million THB. For the price over 12 million THB, the proportion of the houses with 3 to 4 bedrooms are noticeable, if compared with the lower price houses as shown in Fig 3(c). Besides, there are 5 bedrooms houses within the price between 13 to 17 million THB.

In Fig. 4, we present the relationship between attribute 'SalePrice' and other attributes, namely 'Size' and 'Roads'. It can be seen that all houses with price over 10 millions THB

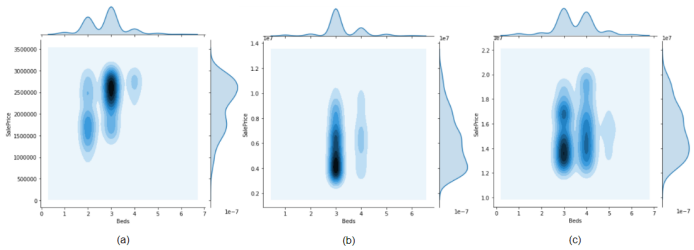

Fig. 3. (a) PDF of attribute 'Beds' over the price less than 3 million THB, (b) PDF attribute 'Beds' over the price ranges between 3 to 12 million THB, and (c) PDF of attribute 'Beds' over price ranges 1.2 to 20 million THB.

TABLE I
ATTRIBUTE OF THE HOUSING ADVERTISEMENTS DATASET

Features	Description	Features	Description
LotSize	Lot size in square feet	DistrictID	ID of district
Baths	Number of bathroom	Size	House size in square metre
Beds	Number of bedroom	Parkings	Number of parking
AvgMinSalePrice	An average price of neighboring houses within 5 km	AvgDistrict	An average price of housing in each district
Lat	Latitude	Lng	Longitude
YearBuilt	Original construction date	ListingID	Housing ID (in website)
SalePrice*	Housing price		
MiniShop†	Number of convenience store within 1 km.	PubTran†	Number of public transport station within 1 km.
Starbucks†	Number of coffee: Starbucks within 1 km.	Roads†	Number of roads within 1 km.

*The target data.
†The data come from OSM.

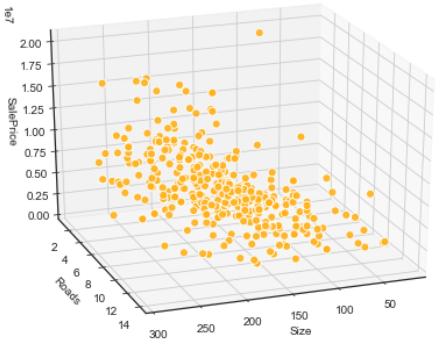

Fig. 4. The relationship between attributes 'SalePrice', 'Roads' and 'Size'.

has at least 100 square meters for a minimum size. Moreover, the houses that connect around 10 roads have a possibility of having price lower than 2.5 millions. While their sizes are between 20 to 160 square meters. In contrast, the houses with high price (Ex., more than 15 million THB) have high possibility to connect with the few number of the road (Ex., 2).

We further investigate the relationship between attributes 'SalePrice' and 'Size' as well as 'SalePrice' and 'PubTran'.

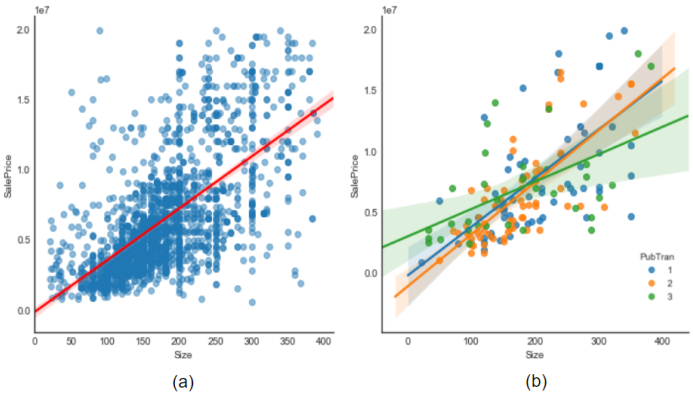

Fig. 5. (a) the relationship between attributes 'SalePrice' with 'Size', (b) the relationship between attribute 'SalePrice' with attributes 'Size' and 'PubTran'.

As shown in Fig 5(a), It is obvious that attribute 'Size' has strong relationship with 'SalePrice'. Note that the houses with larger size have shown potential to be more expensive than the houses with smaller size. In Fig 5(b), It can be also seen that the houses with larger size which connect one more public transport stations tend to have high price.

From Fig. 2 to Fig. 5, we discover that attributes including 'Beds', 'Baths', 'LotSize', 'Size', 'AvgDistrict', 'AvgMinSalePrice', 'PubTran' and 'Roads' can be utilized as the features for our model which will be describe in section III.

III. METHODOLOGY

A. The stacking CNN and various ensemble ML models

This section, we describe the procedure of stacking models which utilize image features to improve the predictive accuracy. Initially, we obtained image features from CNN, VGG-19 network [12], along with fine-tuned weights from ImageNet database from http://www.image-net.org. The image features will be used with the selected features from dataset in Section II(B) as the input of the ensemble ML model. Here, we utilize three remarkable ensemble ML classifiers, namely RF, XGBoost and AdaBoost.

- We first normalize the image pixel values from [0, 255] to [0, 1] to make VGG-19 network easier to train.
- Remove the top layer of VGG-19 network, then add five neuron nodes and one output softmax, which can classify images into two categories: low price (lower than 3 million THB) and high price (more than 7 million THB)
- Feed the normalized image vectors into the modified VGG-19 network.
- After training, the weights from those additional five neuron nodes will be used as the features together with the selected features from the database.
- Feed all features to ML classifier.
- Receive the results.

In our study, we specifically the detail of our data preparation as follows. We first split dataset randomly into 3,578 (80%) training set and 894 (20%) test set. The training set is utilized for training and test set is utilized for the validation. Missing values are replaced with mean value of the corresponding feature. For images, we have 3479 images

in our dataset, including both interior and exterior views. In VGG-19 network, the image training set had been trained with 10 epochs for performing feature extraction.

B. The stacking model with linear regression

In order to further improve the predictive accuracy, we herein propose the concept of collaborative between linear regression technique and a stacking model. Unlike the conventional stacking model, the proposed model utilizes a linear regression technique to average between the actual and the predicted values.

In the proposed model, the results from a stacking model are categorized into 5 sections: Section 1, where the price is lower than 3 million THB, Section 2, where the price ranges between 3 to 6 million THB, Section 3, where the price ranges between 6 to 10 million THB, Section 4, where the price ranges between 10 to 20 million THB and finally Section 5 is the price higher than 20 million THB. Furthers, the best-fitting straight graph is constructed through an associate (P_t, A_t) point in each section, where P_t is denoted as the predicted price and A_t is denoted as the actual price.

The linear regression equation in each section is given by

$$lr_i = b + mP_t \tag{1}$$

where b is denoted as an intercept and m is denoted as a slope in the fitted graph. lr_i is the value lie on the graph of section i, where $i \in \{1, 2, 3, 4, 5\}$. Finally, the final prediction in each section can be estimated through (1). The state diagram of our proposed stacking model with linear regression model is presented in Fig. 6.

IV. Results

This section, we demonstrate our results in order to characterize the proposed model and compare it with its individual ensemble and stacking models. In this paper, MAPE is used to measure the predictive accuracy of the models.

MAPE is given by

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \tag{2}$$

where n is the number of of fitted data. Obviously, MAPE provides an average of percentage errors for the model.

In Table II, MAPE of ensemble models, including RF, XGBoost and AdaBoost with and without CNN are presented. Without CNN, AdaBoost performs better than RF and XGBoost. However, its performances are unacceptable. When integrating with CNN, CNN-RF can reduce an error from its individual model by 113.03 %. CNN-XGBoost and CNN-AdaBoost can reduce an error from their individual models by 60.54% and 10.60%, respectively. It is obvious that the most effective stacking model is CNN-XGBoost, which can provide better performance than that CNN-RF and AdaBoost.

TABLE II
MAPE of Random Forest, XGBoost and AdaBoost classifiers with and with out CNN.

	RF	XGBoost	AdaBoost
without CNN	153.64 %	99.92 %	72.48 %
with CNN	40.61 %	39.38 %	61.88 %

TABLE III
MAPE of stacking CNN-Random Forest, CNN-XGBoost and CNN-AdaBoost using linear regression model.

	CNN-RF with linear regression	CNN-XGBoost with linear regression	CNN-AdaBoost with linear regression
Section 1	24.97 %	24.78 %	24.83 %
Section 2	16.28 %	16.36 %	16.80 %
Section 3	12.16 %	12.86 %	13.78 %
Section 4	15.35 %	15.18 %	15.21 %
Section 5	25.20 %	27.64 %	29.41 %
Total	17.83 %	18.20 %	18.74 %

To investigate the proposed models, Table III shows the MAPE of stacking CNN-RF, CNN-XGBoost and CNN-AdaBoost models, which integrated with linear regression technique. In Table III, CNN-random forest with linear regression technique provides the best performance in total. It can reduce an error to 22.78 %. While CNN-XGBoost and CNN-AdaBoost with linear regression can reduce an error to 21.78 % and 43.14 %, respectively.

Conclusion

We proposed a stacking model for predicting the price of the 2^{nd} hand houses in Thailand. The original dataset was obtained from a leading Thai real estate website and OSM which includes images, intrinsic and extrinsic attributes. We explored and visualized the dataset to obtain insights and examined the key features affecting the price. According to our data analysis, we discovered that some features including 'Beds', 'Baths', 'LotSize', and 'Size' have strong correlation to the price for our dataset. Additionally, we included the average price of neighboring houses within 5 km. as well as the number of road and public transport station within 5 km. around the house into the set of feature.

In order to identify the most effective stacking model for our dataset, we evaluated the performances of RF, XGBoost and AdaBoost, which are powerful ensemble-based models. Our experimental results demonstrated that AdaBoost model can provide a better performance in terms of the MAPE. Furthermore, we found that using image features from deep learning can lead to improved performance. The results demonstrated that the stacking CNN-XGBoost model outperformed other individual techniques. Finally, linear regression can be used to improve the final results. Although obtaining an acceptable error level for industry usage is challenging, our work demonstrates that the combination of stacking image features extracted by CNN and refining the prediction from the ensemble model by a simple linear regression can reduce the error over 80%.

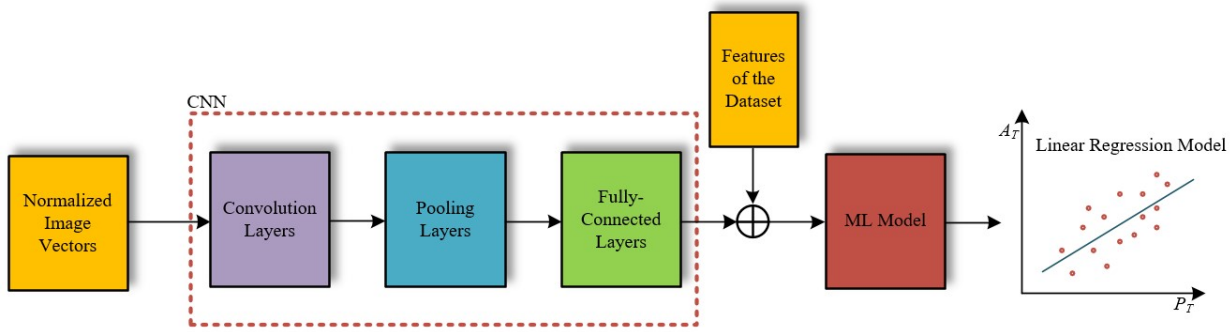

Fig. 6. Architecture of the proposed model.

ACKNOWLEDGMENT

This research was supported by the joint research project between Home Dot Tech Co., Ltd. and the Faculty of Engineering, Chulalongkorn University. We would like to express a special thanks to Mr. Tanawat Promtanapat for early explorations done on this work. PNP was partially supported by "Grants for Development of New Faculty Staff, Ratchadaphiseksomphot Endowment Fund, Chulalongkorn University".

REFERENCES

[1] Q. Truong, M. Nguyen, H. Dang, and B. Mei. "Housing Price Prediction via Improved Machine Learning Techniques." Procedia Computer Science 174 (2020): 433-442.

[2] J. Mu, F. Wu, and A. Zhang. "Housing value forecasting based on machine learning methods." Abstract and Applied Analysis. Vol. 2014. Hindawi, 2014.

[3] T. Li, J, Leng, L. Kong and et al. DCNR: deep cube CNN with random forest for hyperspectral image classification. Multimed Tools Appl 78, 3411–3433 (2019).

[4] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2998-3000.

[5] S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017, pp. 319-323.

[6] S. Fadil, Symes, w. William. (1986). "Linear inversion of band-limited reflection seismograms". SIAM Journal on Scientific and Statistical Computing. SIAM. 7 (4): 1307–1330.

[7] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia, 2018, pp. 35-42.

[8] C. Corinna, V. Vladimir. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297.

[9] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 1936-1939.

[10] R. Xudie, G. Haonan, L. Shenghong, W. Shilin and L. Jianhua. (2017). A Novel Image Classification Method with CNN-XGBoost Model. In: C. Kraetzer, YQ Shi, J Dittmann, H. Kim. (eds) Digital Forensics and Watermarking. IWDW 2017.

[11] C. Spearman, (1904). The proof and measurement of correlation between two things, American Journal of Psychology 15, 72–101

[12] L. Yang, Y. Zhang, Y. Xu, J. Wang, and Z. Miao. "Robust scale adaptive kernel correlation filter tracker with hierarchical convolutional features." IEEE Signal Processing Letters 23, no. 8 (2016): 1136-1140.