House Price Prediction Using Machine Learning Via Data Analysis

Vachana J Rai[1], Sharath H M[2], Sankalp M R[3], Sanjana S[4], Bhavya Balakrishnan[5]

[1,2,3,4] Students, Department of Computer Science & Engineering, T John Institute of Technology, Bengaluru, India
[5]Asst. Professor, Department of Computer Science & Engineering, T John Institute of Technology, Bengaluru, India

Abstract - *Research teams are increasingly adopting machine learning models to execute relevant procedures in the field of house price prediction. As some research did not take into account all available facts, influencing house price forecast and produces inaccurate results. The House Price Index (HPI) is a popular tool for estimating changes in house costs depending on factors such as location, population, industrial growth, and economic prospects. This paper gives a general overview of how to anticipate price of houses based on customer requirements utilizing traditional data and advanced machine learning models, together with regression techniques and python libraries. The effectiveness of our analysis is confirmed by the usage of ANN (Artificial Neural Network), locational attributes, structural attributes, and data-mining's capacity to extract knowledge from unstructured data. This housing price forecast model for Tier-1 cities, with an accuracy of more than 85%, offers enormous benefits, particularly to buyers, developers, and researchers, as prices continue to fluctuate.*

Key Words: **House Price Prediction, Regression Models, Data Analysis, Machine Learning using Python, Algorithms.**

1. INTRODUCTION

As we all know, a house is a basic human need, and costs for them vary from place to place depending on amenities like parking and neighbourhood. The [2] housing markets have a favourable effect on a nation's currency, which is a significant factor in the national economy. Every year, there is a growth in the demand for homes, which indirectly drives up home prices. The issue arises when there are many factors, such as location and property demand, that could affect the house price. As a result, the majority of stakeholders, including buyers, developers, home builders, and the real estate industry, would like to know the precise characteristics or the accurate factors influencing the house price to assist investors in their decisions and to assist home builders in setting the house price. Viewing [5] transaction records may help buyers determine whether they were given a fair price for a home and sellers determine the price at which a home can be sold in a certain area. Finding a workable prediction method is therefore increasingly crucial because the employment of a single classifier is limited.

1.1 Motivation

Many people, whether wealthy or middle class, are concerned about house prices. It is possible to establish a mechanism to predict correct prices based on previous buy/sell values because one can never appraise or estimate the pricing of a house based on the neighborhood or amenities offered. Making sure everyone can purchase a home at the best price is the key goal.

1.2 Objectives

This project is being put forth [1] to forecast house prices and to acquire better and accurate results. It will use a variety of regression methods to provide the most precise and accurate findings. Python programming language is employed for machine learning in order to complete this operation. To determine which regression method produces the most precise and accurate results, the stacking algorithm is performed to multiple regression algorithms.

1.3 Problem Statement

There is no proper price fixation because prices in tier-1 cities are always changing and dependent on a variety of factors, including location, population, and potential future projects. Purchasers, developers, and researchers are all impacted by this. Users need a suitable platform to obtain accurate prices based on historical selling price patterns, employing a variety of features and gathering information from numerous tier-1 city locations.

1.4 Machine Learning using Python

Python is a sophisticated, widely used programming language. In 1991, "GUIDO VAN ROSSUM" invented it. Numerous libraries, including pandas, numpy, SciPy, matplotlib, etc., are supported by Python. It supports Xlsx, Writer, and X1Rd, among other packages. Complex science is performed extremely effectively using it. There are numerous functional Python frameworks.

Machine learning is a branch of artificial intelligence that allows computer frameworks to pick up new skills and enhance their performance with the help of data. It is employed to research the development of computer-based algorithms for making predictions about data. Providing data is the first step in the machine learning process, after which the computers are trained by using a variety of algorithms to create machine learning models. Software engineering's branch of machine learning has significantly altered how people analyse data.

2. RELATED WORKS

According to research [4], the House Price Index (HPI) is used in many nations to gauge variations in the cost of residential real estate, to forecast the average price per square foot of each house, the dataset including a large number of data points and a wide range of factors indicating the house values traded in previous years was employed. To tidy up the data, specify minimal values for the parameters "area" and "price." Prior to creating a regression model, exploratory data analysis is crucial. Researchers are able to identify implicit patterns in the data in this way, which helps them select the best machine learning techniques.

Based on [6] Multiple Linear Regression, the methodology [MLR] The most prevalent type of linear regression is multiple linear regression. With the use of Equations 1 and 2, the [MLR]multiple linear regression is employed as a forecast predictor to demonstrate the relationship between a continuous dependent variable and one or more independent variables: E(Y | X) = 1 + 1X1 + + pXp -(1) Yj = 1+ 1Xj,1 + + pXj,p + j - (2). The various regression techniques might include [LR] One form of linear regression that makes advantage of shrinking is the lasso regression. [ER] In this method, the regularization issue is resolved via elastic net regression. For a non-negative, precisely between 0 and 1, and a 1. The Gradient Boosting algorithm (GBR) is a machine learning method for resolving issues with regression and classification. As a result, a prediction model is created that combines all of the weak prediction models, mostly decision trees. AdaBoosting Regression (AR) is a regression approach designed to combine "simple" and "weak" classifiers to create a "strong" classifier.

The [10] CNN-based prediction model for home prices is a type of feed-forward neural network with a deep structure and convolution processing, which is one of the representative deep learning algorithms. In the experimental phase, the CNN model is constructed using the TensorFlow framework. The activation function for the model's two convolutional layers is the Relu function, and the dropout method is also used to prevent over-fitting. Results from this CNN model experiment is 98.68% accurate. Although [11] the accuracy, precision, specificity, and sensitivity are the four criteria used to assess the success of machine learning systems. The two discrete values 0 and 1 are regarded as distinct classes in the work. If the class value is 0, we assume that the house's price has reduced, and if the class value is 1, we assume that the house's price has grown. The accuracy levels attained with this technique for various machine learning methods are as follows: Random Forest - 86%, Support Vector Machine - 82%, and Artificial Neural Network - 80%.

As in [12] one of the values of the tax object (NJOP) of the land and the value of the tax object of the building are the factors that can be quantitatively calculated to determine the selling price of the home. Numerous elements, like the building's age and strategic placement, have an impact on both criteria. To acquire the best prediction accuracy, the initial house price prediction is difficult and calls for the finest methodology. Fuzzy logic becomes one of the strategies that can be employed in solving the problem of estimating the sale price of a house that has an uncertainty parameter. For predicting residential property prices, predictions can also make use of the K-Nearest Neighbors method in addition to fuzzy logic and artificial neural networks.

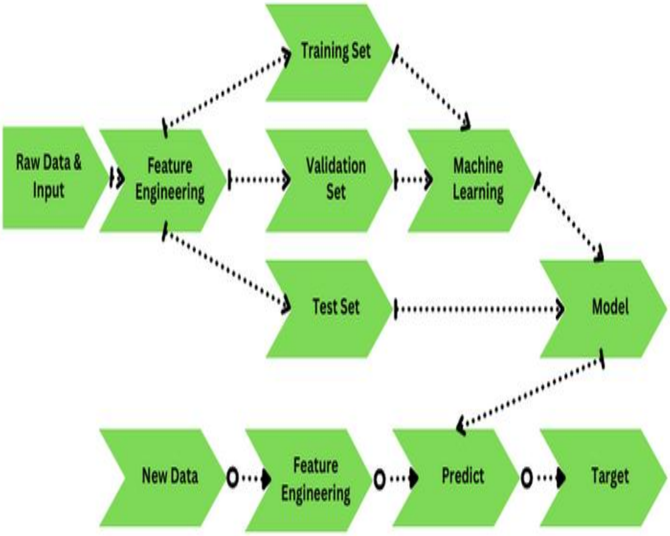
In [13] the nonlinear model for real estate's price variation prediction of any city, based on leading and concurrent economic indices, is established using two techniques, similar to those used in back propagation neural network (BPN) and radial basis function neural network (RBF). The outcomes of the predictions are contrasted with the Public House Price Index. The two indices of the price fluctuation that are chosen as the performance index are the mean absolute error and root mean square error.

Similar to [14], the GA-BP model, which combines the genetic algorithm (GA) and back propagation neural network (BPNN), was used to predict the housing price along the urban rail transit line. This study examined the relationship between accessibility along the urban rail transit line and the change in housing price. By using the cost of residential areas near a busy metro line as an example and contrasting the performance of the GA-BP model and the BP model, the model's dependability was confirmed. It was discovered that the mean square error of the prediction outcomes is much lower and that the average absolute error rate of the GA-BP model is 6.91%, which is 6.45% lower than that of the BP model. As can be shown, the GA-BP model-based price prediction algorithm for those residential areas near the urban line is accurate and completely exploits the potential relationship between the affordability of housing and the accessibility of the urban transportation network.

As with reference to [15], an unsupervised learnable neuron model (DNM) by including the nonlinear interactions between excitation and inhibition on dendrites is used. We fit the House Price Index (HPI) data using DNM, after which we project the trends in the Chinese housing market. We compare the performance of the DNM to a conventional statistical model, the exponential smoothing (ES) model, in order to confirm the efficacy of the DNM. The accuracy of the two models' forecasts is assessed using three quantitative statistical metrics: normalized mean square error, absolute percentage of error, and correlation coefficient. According to experimental findings, the suggested DNM outperforms the ES in each of the three quantitative statistical criteria.

3. IMPLEMENTATION AND WORKING

3.1 Architectural Diagram

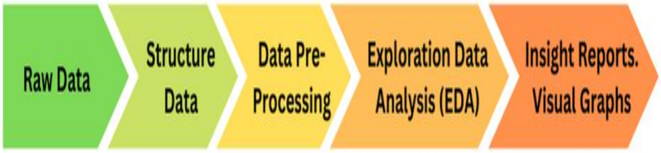


The [7] the basic structural working methodology is based on the above flowchart.

3.2 Data Collection and Data Cleaning

There [3] are several methods and procedures used in data processing. Data from real estate properties in Tier-1 cities was gathered from several real estate websites. The information would include attributes like location, carpet size, built-up area, property age, zip code, etc. We need to gather structured and organized quantitative data. Before beginning any machine learning research, data must be collected. Without dataset validity, there would be no sense in evaluating the data.

Errors are found and removed throughout the data cleaning process in order to maximize the value of the data. To guarantee that accurate and correct information is available from the record set, table, or dataset, it replaces untidy information. The main goal of data cleaning is to provide a dynamic estimation of information


Raw Data
Structure Data
Data Pre-Processing
Exploration Data Analysis (EDA)
Insight Reports. Visual Graphs

3.3 Data Preprocessing and Analysis of Data

The [8] dataset needs to be pre-processed before using models to predict house prices. First, a missing data investigation is carried out. A number of missing patterns are systematically evaluated since they are crucial in determining the best ways to handle missing data. Since it is challenging to impute these missing values with an acceptable level of accuracy, columns with more than 55% of their values missing are eliminated from the original dataset. Additionally, the result variable's values are missing from a large number of rows (Price). The observations with missing values in the Price column are eliminated since the imputation of these values can enhance bias in the input data.

3.4 Data Visualization

Due to its capacity to successfully investigate challenging concepts and find novel patterns, data visualization is used in many high-quality visual representations.

3.5 Cross Validation and Training the Model

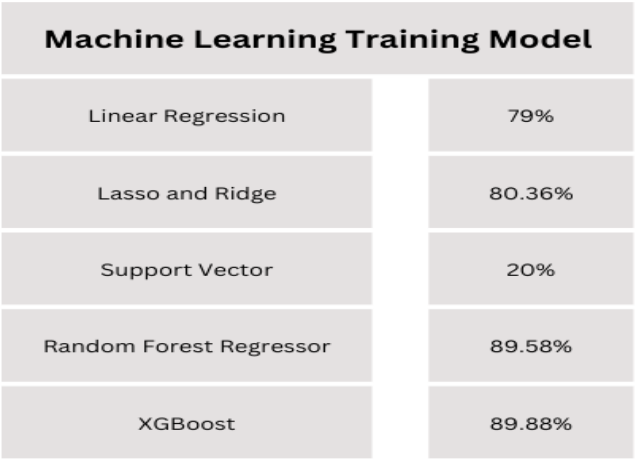
In order to train a machine learning model using a subset of the dataset, cross validation is performed. Training is important to obtain accuracy when dividing data sets into "N" sets for assessments of the model that has been constructed.

We must first train the model because the data is divided into two modules: a test-set and a training-set. The target variable is a part of the training set. The training data set is subjected to the decision tree regressor method. Using a tree-like structure, the decision tree creates a regression model.

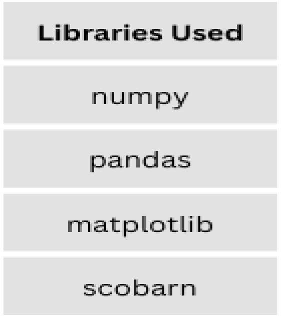
3.6 Testing and Integration with UI

A web framework Flask, gives you the technology, tools, and libraries necessary to create a web application. Flask is a framework primarily used for integrating Python models because it is simple to build together routes.

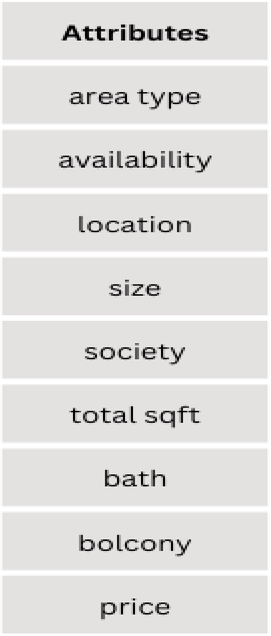
House prices are forecasted using the training model and a test dataset. The front end is then connected with the trained model using Python's Flask framework. After creating the model and successfully producing the desired result, the integration with the user interface (UI) is the next stage, and flask is employed for this.



Machine Learning Training Model	
Linear Regression	79%
Lasso and Ridge	80.36%
Support Vector	20%
Random Forest Regressor	89.58%
XGBoost	89.88%



Libraries Used
numpy
pandas
matplotlib
scobarn

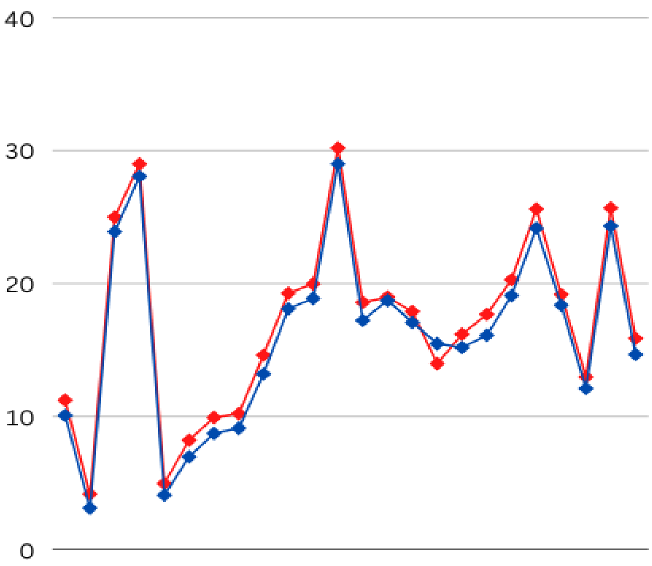


Attributes
area type
availability
location
size
society
total sqft
bath
bolcony
price

4. RESULTS

Python-based data mining techniques are used to produce the desired results. Numerous variables that have an impact on home pricing are taken into consideration and further developed. The idea of using machine learning to carry out the desired task has been considered. Data collection is started first. Then, data cleaning is carried out to make the data clean and error-free. The next step is data pre-processing. The distribution of data in various forms is then intended to be depicted through the creation of various plots using data visualization. In the end, the business costs of the homes were calculated precisely. In addition to employing regression techniques, several classification algorithms are taken into account and used on our house pricing dataset, including the SVM algorithm, decision tree algorithm, Random Forest classifier, etc. a way that would assist people in purchasing homes at a price that is affordable and within their means. Various algorithms are used to determine the homes' sales prices. The sales prices have been determined more precisely and accurately. The public would benefit greatly from this. Various data mining methods are used in Python to produce these findings. It is important to think about and address the numerous aspects that have an impact on home pricing. We received help from machine learning to finish our assignment. Data collecting is done first.

Then, data cleaning is done to make the data clean and remove all of the errors. The data pre-processing is then completed. Then, several graphs are made using data visualization. This has shown how data is distributed in several ways. Furthermore, the model is prepared for use and tested. Some categorization techniques were discovered to have been used on our dataset, whereas others had not. Therefore, the algorithms that were not being used on our house pricing dataset were removed, and efforts were made to increase the accuracy and precision of the algorithms that were. A unique stacking approach is suggested in order to increase the precision of our classification systems. In order to get better outcomes, it is crucial to increase the algorithms' accuracy and precision. The people would not be able to estimate the sales prices of houses if the results were inaccurate. Data visualization was also used to improve accuracy and outcomes. Various algorithms are used to determine the homes' sales prices. The sales prices have been determined more precisely and accurately. The public would benefit greatly from this.



5. FUTURE SCOPE

The future objective is to expand the dataset to other Indian states and cities, as it currently only comprises Tier-1 cities. We'll be integrating map into the system to make it even more informative and user-friendly. As [20] numerous important factors influence property values. It is a good idea to add extra elements if statistics are available, such as income, salary, population, local amenities, cost of living, annual property tax, school, crime, and marketing information. In the near future, we'll give a comparison of the price projected by the system and the price from real estate websites like Housing.com, magicbricks.com etc., for the same user input. We will also suggest real estate properties to the consumer depending on the anticipated pricing to further simplify things for them.

6. CONCLUSION

This work [16] uses various machine learning algorithms to undertake an analytical analysis of the effects of real estate market fluctuations on real estate health and trends. The ability to estimate house prices is crucial for minimising the effects of property valuation and economic expansion in complicated real estate systems. Among the several machine learning methods, XGBoost is applied to enhance home price prediction in intricate real estate systems. The [17] trial and error is used to identify the best ANN model. But [18] ANN has its drawbacks. The inconsistent nature of the findings is one major issue. The ANN model is capable of self-learning during the training phase and modifying the weights appropriately to reduce the error. As a result, it is impossible to tell whether the model produced the best result. A series of tests were run on a public real estate dataset in order to assess and contrast the suggested model. According to the experimental data, the XGBoost model is more successful than other baseline machine learning prediction methods at predicting home prices for real estate, attaining 89% of the measure. Both market players and banks can use this result. As [19] the market players are constantly curious about the likelihood of their homes selling. However, if a borrower is unable to repay the loan, it would be fascinating for the banks to know how much chance they have of selling the homes they have taken as collateral. The maintenance of financial stability depends heavily on the latter.

7. ACKNOWLEDGEMENT

We thank, Dr. Thomas P John (Chairman), Dr. Suresh Venugopal P (Principal), Dr Srinivasa H P (Vice-principal), Ms. Suma R (HOD – CSE Department), Dr. John T Mesia Dhas (Associate Professor & Project Coordinator), Ms. Bhavya Balakrishnan (Assistant Professor & Project Guide), Teaching & Non-Teaching Staffs of T. John Institute of Technology, Bengaluru – 560083.

REFERENCES

[1] Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla, “Prediction of House Pricing Using Machine Learning with Python,” IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4, DOI: 10.1109/ICESC48915.2020.9155839

[2] Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, Ismail Ibrahim, “House Price Prediction using a Machine Learning Model: A Survey of Literature,” I.J. Modern Education and Computer Science, 2020, 6, 46-54, DOI: 10.5815/ijmecs.2020.06.04

[3] Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik, Shila Jawale (Guide), “House Price Forecasting Using Machine Learning,” Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020 at: https://ssrn.com/abstract=3565512

[4] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019) - Procedia Computer Science 174 (2020) 433–442, DOI:10.5815/ijmecs.2020.06.04

[5] Pei-Ying Wang, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang, Szu-Hao Huang, “Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along with Joint Self-Attention Mechanism,” IEEEAccess – DOI: 10.1109/ACCESS.2021.3071306

[6] CH. Raga Madhuri, Anuradha G, M.Vani Pujitha, “House Price Prediction Using Regression Techniques: A Comparative Study,” IEEE 6th International Conference on smart structures and systems ICSSS 2019, DOI: 10.11098/ICSSS.2019.8882834

[7] P. Durganjali, M. Vani Pujitha, “House Resale Price Prediction Using Classification Algorithms,” IEEE 6th International Conference on smart structures and systems ICSSS 2019 - 978-1-7281-0027-2/19/$31.00 ©2019 IEEE, DOI: 10.1109/ICSSS.2019.8882842

[8] The Danh Phan, “Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia,” 2018 International Conference on Machine Learning and Data Engineering (iCMLDE) - 978-1-7281-0404-1/19/$31.00 ©2019 IEEE, DOI 10.1109/iCMLDE.2018.00017

[9] Yajuan Tang, Shuang Qiu, Pengcheng Gui, “Predicting Housing Price Based on Ensemble Learning Algorithm,” DOI: 10.1109/IDAP.2018.8620781

[10] Yong Pia, Ansheng Chen, Zhendong Shang, “Housing Price Prediction Based on CNN,” 9th International Conference on Information Science and Technology (ICIST), Hulunbuir, Inner Mongolia, China, DOI: 10.1109/ICIST.2019.8836731

[11] Debanjan Banerjee, Suchibrota Dutta, “Predicting the Housing Price Direction using Machine Learning Techniques,” IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017) - 978-1-5386-0814-2/17/$31.00 ©2017 IEEE, https://doi.org/10.1109/ICPCSI.2017.8392275

[12] Muhammad Fahmi Mukhlishin, Ragil Saputra, Adi Wibowo, "Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbour," 2017 1st International Conference on Informatics and Computational Sciences (ICICoS) - 978-1-5386-0903-3/17/$31.00 © 2017 IEEE, DOI: https://doi.org/10.1109/ICICOS.2017.8276357

[13] Li Li and Kai-Hsuan Chu, "Prediction of Real Estate Price Variation Based on Economic Parameters," Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017 - Meen, Prior & Lam (Eds) - ISBN 978-1-5090-4897-7, DOI: https://doi.org/10.1109/ICASI.2017.7988353

[14] Li Ruo-qi and Hu Jun-hong, "Prediction of Housing Price Along the Urban Rail Transit Line Based On GA-BP Model and Accessibility," 2020 IEEE 5th International Conference on Intelligent Transportation Engineering - 978-1-7281-9409-7/20/$31.00 ©2020 IEEE, DOI: https://doi.org/10.1109/ICITE50838.2020.9231460

[15] Ying Yu, Shuangbao Song, Tianle Zhou, Hanaki Yachi, Shangce Gao, "Forecasting House Price Index of China Using Dendritic Neuron Model," 978-1-5090-3484-0/16/$31.00 ©2016 IEEE, DOI: https://doi.org/10.1109/PIC.2016.7949463

[16] Bandar Almaslukh, "A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems," 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), DOI: 10.1109/TAAI51410.2020.00047

[17] Wan Teng Lim, Lipo Wang, Yaoli Wang, and Qing Chang, "Housing Price Prediction Using Neural Networks," 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) - 978-1-5090-4093-3/16/$31.00 ©2016 IEEE, DOI: 10.1109/FSKD.2016.7603227

[18] Lipo Wang, Fung Foong Chan, Yaoli Wang, and Qing Chang, "Predicting Public Housing Prices Using Delayed Neural Networks," 2016 IEEE Region 10 Conference (TENCON) — Proceedings of the International Conference - 978-1-5090-2597-8/16/$31.00 ©2016 IEEE, DOI: 10.1109/TENCON.2016.7848726

[19] Ceyhun Abbasov, "The prediction of the chance of selling of houses as the factor of financial stability," 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), DOI: 10.1109/icaict.2016.7991786

[20] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Prices Prediction," Proceedings of the 2017 IEEE IEEM - 978-1-5386-0948-4/17/$31.00 ©2017 IEEE, DOI: 10.1109/IEEN.2017.8289904