Grenze International Journal of Engineering and Technology, Jan Issue

Estimating Bangalore Housing Prices using Textual Attributes and Image Data

Dr. Seema Kolkur[1], Ishan Saksena[2], Shashwat Satao[3] and Darsh Shah[4]
[1-4]Thadomal Shahani Engineering College/Computer Engineering Department, Mumbai, India
Email: seema.kolkur@thadomal.org, saksenaishan@gmail.com, sataoshashwat0@gmail.com, darsh0908@gmail.com

***Abstract*—The majority of current automated algorithms for estimating house prices only take into account textual information like the home's neighbourhood and the number of rooms. A human representative makes a physical inspection of the home and makes an estimate of the actual cost. In this work, we suggest merging written information about the property with visual elements extracted from images of the house. The combined features are fed into a Convolution Neural Network (NN), which produces a single output that is an estimate of the price of the home. For the purpose of developing and evaluating our network, we have gathered a housing dataset that includes both visual and textual features. The dataset consists of the price and other features of 500 homes from Bangalore along with images of those homes.**

***Index Terms*— Housing Price, Bangalore, CNN, SVR, textual data, and image data.**

I. Introduction

The housing market is a key factor in how the economy is shaped. By raising the pace of home sales, employment, spending, housing rehabilitation and construction stimulates the economy. It also influences the demand for goods from other important businesses, such those that provide building
materials and durable goods for the home (Li et al., 2011). Recent studies demonstrate that the profitability of financial institutions is impacted by the housing market, which in turn has an impact on the surrounding financial system.
Additionally, the housing industry serves as a crucial indication of the real sector of the economy as well as asset values that aid in forecasting inflation and production (Li et al., 2011). The arduous conventional price prediction technique is based on the cost and sales price comparison, which are inaccurate and do not have an established standard or certification procedure (Khamis and Kamarudin, 2014).
In order to better establish policies, prevent inflation, and aid individuals in making good investment decisions, it is necessary to have an accurate automated prediction of housing price (Li et al., 2011). Predicting home values is a very difficult task due to the illiquidity and heterogeneity in both the physical and geographical views of the housing market.
Additionally, a delicate interplay between the home price and a few other macroeconomic variables makes the forecast technique extremely difficult. Previous research was done to identify the most crucial elements influencing housing prices. All of the earlier studies focused on the homes' linguistic characteristics (Khamis and Kamarudin, 2014; Ng and Deisenroth, 2015; Park and Bae, 2015). So, in order to leverage both visual and verbal qualities in the price assessment process, we decided to integrate them. According to (Limsombunc et al., 2004), the neighborhood, size, and number of bedrooms and bathrooms all have an impact on a home's pricing. The cost of a home increases with the number of bedrooms and bathrooms.

Grenze ID: 01.GIJET.9.1.713

II. RELATED WORK

Some effort has been done to automate the real estate price appraisal process during the past ten years. Successes included highlighting the property's features, including its location, quality, environment, and site. When comparing various techniques, we discovered that they may be divided into two major categories: data aggregation-based models and data disaggregation-based models. Similar to the Hedonic Price Theory, the data disaggregation based models attempt to forecast the house's price based solely on each feature. To predict the price of a house, data aggregation techniques like regression and neural networks must take into account all of the property's characteristics.

The Hedonic Price is a prime example of a Data Disaggregation Model. Theory in which the characteristics of real estate determine its price. A collection of implicit pricing is defined by the real estate's related qualities. The marginal implicit values of the qualities are created by differentiating the hedonic pricing function with respect to each characteristic (Limsombunc et al., 2004). The issue with this approach is that it ignores the variations across properties within the same geographic region. It is therefore seen as being unrealistic. By contrasting the outcomes of data aggregation and disaggregation, Flitcher et al. attempted to determine the optimal method for estimating the property price in (Fletcher et al., 2000).

They discovered that the outcomes of agglomeration are more precise. They also discovered that some coefficients' hedonic prices are not constant since they vary with location, age and property type. They thus came to the realization that while analyzing these changes, hedonic analysis can be useful, it cannot be used to estimate the price based solely on a characteristic. They also learned that the property's location has a significant impact on its pricing. Neural Networks are the most popular model for the Data Aggregation model.

In (Khamis and Kamarudin, 2014), Bin Khamis examined the effectiveness of the Neural Network and Multiple-Linear Regression (MLR). NN outperformed MLR in terms of MSE and R2, respectively. The neural network model surpasses the Hedonic model when results from the two models are compared because it achieves a better R2 value by 40.769% and a lower MSE by 45.4314%. The Hedonic model's information deficit might be what's behind its subpar performance. The Neural Network Model does have some limits, though, as the predicted price is not the exact price but rather a near approximation.

This is as a result of the challenge in getting accurate market data. Additionally, the temporal impact, which neural networks cannot automatically manage, is crucial to the estimating process. This suggests that several more economic factors that are challenging to account for in the estimation process also have an impact on the price of real estate. In this article, we'll look into how the estimation process is affected when visual characteristics and textual attributes are combined. The SVR and the NN are the two estimating models that will be explored.

III. DATASET DESCRIPTION

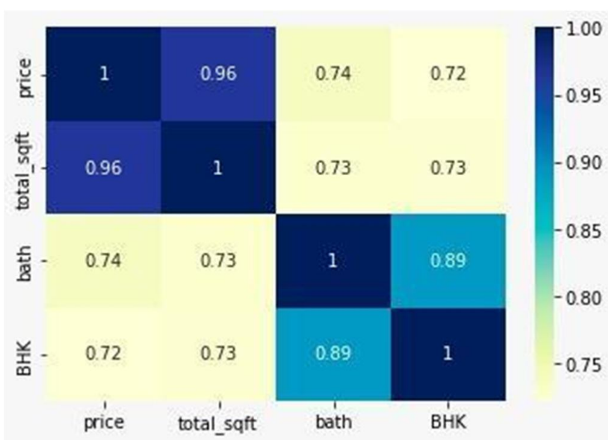

Figure 1. Correlation between Variables

The dataset was created by scraping data from the website housing.com using Beautiful soup. There are two forms of data, textual and image data. Textual data consists of features that describe house-property in Bengaluru. The features can be explained as follows:

1. Price- Value of the property in lakhs.

2. BHK - Number of bedrooms
3. bath - Number of bathrooms.
4. total_sqft - The area of the house in square feet.

TABLE I. DATASET SUMMARY

	area	bath	BHK	price
mean	1689.489	2.620	2.622	110.671
std dev	925.921	0.943069	0.9431	71.878
count	500	500	500	500

In table 1, the mean and standard deviation of all the data points, for features like area, bath, BHK and price are computed.
The image data is a grid of four images: The bedroom, the bathroom, the kitchen and the apartment complex.Fig 2 shows a sample image. From fig 1, the correlation between the different variables in the dataset can be seen. It can be seen that the housing price has a strong positive correlation with the bath (number of bathrooms in the house), BHK(number of bedrooms in the house) and total_sqft(area of the house). Thus, these features were given greater weightage while predicting the house price using the textual data.

Figure 2. Sample Image from Dataset

IV. EVALUATION METRICS

The evaluation metrics which we have used to calculate the accuracy of the models are:

A. R-squared error

In a regression model, the R-squared statistic quantifies the percentage of variance in the dependent variable that can be accounted for by the independent variable(Fig-3). The range of the r-squared model value is 0 to 1, with a greater value indicating better data fit by the model. R-squared measures how well the variation of one variable accounts for the variance of the second, as opposed to correlation, which describes the strength of the relationship between independent and dependent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{predicted} - y_{observed})^2}{\sum_{i=1}^{n} (y_{predicted} - \bar{y}_{observed})^2}$$

Figure 3. Formula for R-Squared error

B. MSE

The mean squared error of an estimator calculates the average of the squares of the errors, or the average squared difference between the estimated values and the actual value. (or a process for estimating an unobserved variable). A measure of an estimator's quality is the MSE. It is always a positive number that gets smaller as the error gets closer to zero since it is derived from the square of the Euclidean distance.

V. PROPOSED BASELINE SYSTEM

A. Support Vector Regression (SVR)

Support vector machines are used for estimating and regressing multidimensional functions. SVMs are based on the idea of the best possible class separation and are derived from statistical learning theory. SVMs use a high dimensional feature space to train and produce prediction functions that are extended on a subset of support vectors.

B. Convolutional Neural Networks (CNN)

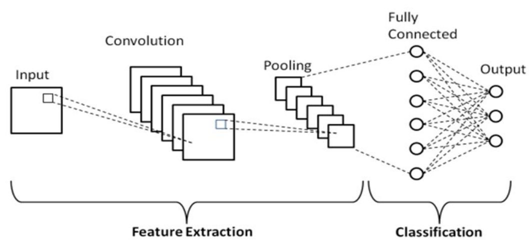

Figure 4. Convolutional Neural Networks

A network or circuit of biological neurons is referred to as a neural network. Therefore, a neural network can either be a biological neural network made up of biological neurons or it can be an artificial neural network created to address problems with artificial intelligence (AI). Artificial neural networks use weights between nodes to resemble biological neuron connections. A positive weight denotes an excitatory connection, whereas a negative weight denotes an inhibitory one. Prior to combining, each input is given a weight. This action is referred to as a linear combination. Finally, the output's amplitude is controlled by an activation function.
As shown in figure 4, a Convolutional Neural Network (ConvNet/CNN) is a Deep Learning method that can take in an input picture, give various elements and objects in the image importance (learnable weights and biases), and be able to distinguish between them. Comparatively speaking, a ConvNet requires substantially less pre-processing than other classification techniques. ConvNets have the capacity to learn these filters and properties, whereas in basic techniques filters are hand-engineered.

VI. EXPERIMENTS AND RESULTS

Our study’s primary objective is to determine the influence of integrating aesthetic characteristics of homes in estimates of their pricing. Hence we utilized Support Vector Regression and Neural Networks Models to investigate the link between the quantity of visual features and the quality of the estimation. SVR uses only textual attributes whereas CNN uses both textual and image data.

A. SVR Experiments

In the SVR model, 428 homes—or 80% of the dataset—were used for training, and 107 homes—or 20% of the dataset—were utilized for testing. The MSE and R values were calculated on the training dataset and testing dataset.

B. Convolution Neural Network Experiments

By experimenting with various counts of neurons in the hidden layer, it was found that having 4 neurons is the ideal architecture. Theneural network uses textual data as well as image data to make its predictions. To balance the number of input and output nodes, the number of hidden nodes was set at four. So the effective architecture is 8(4 textual + 4 images)-4-1 as input-hidden-output neurons respectively. Our neurons had sigmoid activation function and trained using the error-back propagation approach in the Levenberg Marquardt variation. During our tests, this design delivered the greatest outcomes. Our dataset was separated into three sections: training (70%), validation (15%), and testing (15%).The validation and training error start to stabilise around 7 epochs, a measurement of how many times all of the training vectors are used once to update the weights, therefore we terminated the training to prevent over-fitting.

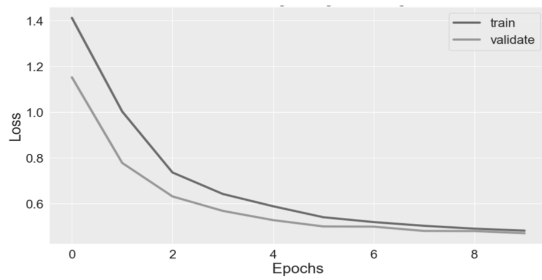

Figure 5. MSE vs Epochs

Figure 5 displays the Network's performance while emphasising the training and validation's MSE loss as well as the time (around 8 epochs) when the training process was halted to prevent over-fitting.

C. Results

TABLE II. EXPERIMENTAL RESULTS

	Training MSE	Training R value	Testing MSE	Testing R value
SVR	1.4 E9	0.903	2.43 E9	0.813
Neural Network	1.2 E9	0.96	1.56 E9	0.956

According to the findings listed in table 2, our model succeeds on the testing set with an MSE of 1.56E9 and an R-Value of 0.956. This means textual attributes along with image data provide better results in predicting house prices.

VII. CONCLUSION

This research uses a dataset for house price prediction with both visual and textual information. It was demonstrated that visual and textual data produced greater estimation accuracy. The relationship between the image features are presently being researched by us.

REFERENCES

[1] The Danh Plan(2018). Housing Price Predictions using machine learning algorithms:The case of Melbourne City, Australia. Retrieved from International Conference on Machine Learning and Data Engineering: https://ieeexplore.ieee.org/document/8614000

[2] CH Raga Madhuri, Anuradha G, M Vani Puthuja (2019).Housing Price Prediction Using Regression Techniques : A Comparative Study. Retrieved from ICCS: https://ieeexplore.ieee.org/document/8882834

[3] Manasa J, Radha Gupta, Narahari NS (2020). Machine Learning based Predicting House Prices Using Regression Techniques. Retrieved from ICIMIA : https://ieeexplore.ieee.org/document/9074952

[4] Shen Rong, Zhang Bao-wen (2020).The research of regression models in machine learning. Retrieved from IFID: https://www.researchgate.net/publication/326121964_The_research_of_regression_model_in_machine_learning_field

[5] Tri Doan, Jugal Kalita (2015). Selecting Machine Learning Algorithms using Regression Techniques. Retrieved from ICDMW: https://ieeexplore.ieee.org/document/7395848

[6] A. Khamis, N. Kamrudin (2014). Retrieval Comparative Study On Estimate House Price Using Statistical and Neural Network Model. Retrieved from Semantic Scholar: https://www.semanticscholar.org/paper/Comparative-Study-On-Estimate-House-Price-Using-And-Khamis-Kamarudin/e4e60873e45c1b3669c66e57d5b1477c73fcdca2

[7] Visit Limsombunchai, Minsoo Lee (2004). House Price Prediction : Hedonic Price Model vsArtificial Neural Network. Retrieved from Research Gate: https://www.researchgate.net/publication/26398692_House_Price_Prediction_Hedonic_Price_Model_vs_Artificial_Neural_Networ

[8] Li, Y., Leatham, D. J., et al. (2011).Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model. Retrieved from AAEA: https://core.ac.uk/download/pdf/6620741.pdf