Akshay Babu *et al*, International Journal of Computer Science and Mobile Applications,
Vol.7 Issue. 3, March- 2019, pg. 8-15 **ISSN: 2321-8363**
Impact Factor: 5.515

Literature Review on Real Estate Value Prediction Using Machine Learning

Akshay Babu[1], Dr. Anjana S Chandran[2]
[1]*Scholar, SCMS, Cochin, Kerala, India, akshaybabuab95@gmail.com*
[2]*Assistant Professor, SCMS, Cochin, Kerala, India, anjana@scmsgroup.org*

Abstract
The real estate market is one of the most competitive in terms of pricing and same tends to vary significantly based on numerous factors; forecasting property price is an important module in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features multiple regressions or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result.
Keywords: Uses, Advantages, Literature survey.

1. Introduction
The study on land price trend is deemed to be significant to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investors decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. In order to accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data revealed that the prices show a non- linear characteristic. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and high-rise buildings. Investments in Real Estate Industry has grown significantly high over the years and we have noticed a non-uniform pattern in terms of land pricing. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Over the last two decades there have been a large number of empirical studies analysing land prices. Kilpatrick [2] showed the usefulness of time-series regression model which used economic data to provide forecast of Central Business District (CBD) land price in moving market. Wilson et al. [3] studied the residential property market accounts for a substantial proportion of UK economic activity. Valuers estimate property values based on current bid prices. In this paper, the national housing transaction data was trained using Artificial Neural Networks (ANN), which forecasts future trend of the housing market. Mark and John [4] developed a regression model with vacant land sales. The model explained up to 93% of the market values. Wang and Tian [5] used the

wavelet Neural Network (NN) to forecast the real estate price index. This kind of wavelet NN integrated the merit of the wavelet analysis and the tradition NN. It also compared the forecasting result with smoothing method and the NN forecast. Zhangming [6] forecasted the real estate price index by using the Back Propagation (BP) NN. The BPN used the sigmoid function. Tinghao [7] used the Auto Regressive Integrated Moving Average (ARIMA) model and carried the demonstrative analysis on year data from 1998 to 2006. He used the established model to make the forecast to the real estate price index of 2007. A hedonic regression on the price of land suggested that de facto policy differences between political jurisdictions have had a significant effect on land prices between1970 and 1980. Steven and Albert [9] used 46,467 residential properties spanning 1999 - 2005 and demonstrated that using matched pairs that relative to linear hedonic pricing models, ANN generate lower dollar pricing errors, had greater pricing precision out-of-sample, and extrapolate better from more volatile pricing environments. ANN is better suited to hedonic models that utilize large numbers of variables. Sampath kumar and Santhi [10] studied the land price trend of Sowcarpet which is the central part. They developed statistical model using economic factors and predicted that the annual rise in land price would be of 17%. Urmila [11] reported that the past trends were analysed to ascertain the rate of growth or decline and the trends are used in forecasting. Economic parameters might be introduced to formulate more realistic relationship. Some of the other techniques they Mansural Bhuiyan and Mohammad Al Hasan 2016 [12] use is regression, deep learning to learn the nature of models from the previous results (the property/land which were sold off previously which are used as training data).There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

2. Need for Real Estate Value Prediction

- While our nation continues growth trend and the construction industry lags behind demand, prices will continue to rise while interest rates bump upward.
- Securing investment property now with thorough due diligence and watch investment pay off over the next few years.

3. Advantages of Machine Learning Over Real Estate Value Prediction

- Trends and Patterns Are Identified with Ease
- Machine Learning Improves Over Time
- Machine Learning Lets You Adapt Without Human Intervention
- Enables Automation

4. Literature Survey

Machine learning is a form of artificial intelligence which compose available computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadly classified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning.

Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our-self.

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience.

Machine learning has many application's out of which one of the applications is prediction of real estate. The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The study on land price trend is felt important to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investors decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting.
The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned.
The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and high-rise buildings. Investments started pouring in Real estate Industry and there was no uniform pattern in the land price over the years. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors.
Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features multiple regressions or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result.

4.1 Previous works and studies

Over the last two decades there have been a large number of empirical studies analysing land prices. *Kilpatrick* showed the usefulness of time-series regression model which used economic data to provide forecast of Central Business District (CBD) land price in moving market. *Wilson et al* studied the residential property market accounts for a substantial proportion of UK economic activity. Valuers estimate property values based on current bid prices. In this paper, the national housing transaction data was trained using Artificial Neural Networks (ANN), which forecasts future trend of the housing market. *Mark and John* developed a regression model with vacant land sales. The model explained up to 93% of the market values. *Wang and Tian* used the wavelet Neural Network (NN) to forecast the real estate price index. This kind of wavelet NN integrated the merit of the wavelet analysis and the tradition NN. It also compared the forecasting result with smoothing method and the NN forecast.

Zhangming forecasted the real estate price index by using the Back Propagation (BP) NN. The BPN used the sigmoid function. *Tinghao* used the Auto Regressive Integrated Moving Average (ARIMA) model and carried

the demonstrative analysis on year data from 1998 to 2006. He used the established model to make the forecast to the real estate price index of 2007. A hedonic regression on the price of land suggested that de facto policy differences between political jurisdictions have had a significant effect on land prices between1970 and 1980. *Steven and Albert* used 46,467 residential properties spanning 1999 - 2005 and demonstrated that using matched pairs that relative to linear hedonic pricing models, ANN generate lower dollar pricing errors, had greater pricing precision out-of-sample, and extrapolate better from more volatile pricing environments. ANN is better suited to hedonic models that utilize large numbers of variables. *Sampath kumar and Santhi* studied the land price trend of Sowcarpet which is the central part. They developed statistical model using economic factors and predicted that the annual rise in land price would be of 17%.
Urmila reported that the past trends were analysed to ascertain the rate of growth or decline and the trends are used in forecasting. Economic parameters might be introduced to formulate more realistic relationship.
Some of the other techniques they *Mansural Bhuiyan and Mohammad Al Hasan* 2016 use is regression, deep learning to learn the nature of models from the previous results (the property/land which were sold off previously which are used as training data).There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

Multiple Regression Technique - *Sampathkumar*
Regression analysis is widely used for forecasting. Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. If more independent variables are added, it is able to determine an estimating equation that describes the relationship with greater accuracy. Multiple regressions look at each independent variable and test whether it contributes significantly to the way the regression describes the data.

The general multiple regression equation is

$$Y = a + b1x1 + b2x2 + \cdots bnxn$$

Advantages: The estimates of the unknown parameters obtained from linear least squares regression are the optimal. Estimates from a broad class of possible parameter estimates under the usual assumptions are used for process modelling. It uses data very efficiently. Good results can be obtained with relatively small data sets.

Disadvantages: The outputs of regression can lie outside of the range [0,1]. It has limitations in the shapes that linear models can assume over long ranges. The extrapolation properties will be possibly poor. It is very sensitive to outliers. It often gives optimal estimates of the unknown parameters.

Neural Network Technique – *Sampathkumar*
NN is a computational technology from the artificial intelligence discipline whose architecture emulates the network of nerve cells in the human brain. A NN is a parallel distributed information-processing structure consisting of processing elements (PEs) which contains local memory. NN architecture such as a standard BP NN can be developed by using the various indicators as PEs to be investigated upon. The approach presents the application of NN for modelling the land price trend with the support of economic and social factors. NN model is constructed with 13 indicators that are PEs with one bias node as input. All the input values are normalized using the MinMax. The principle behind normalization process

$$\text{Normalized value, } N = \frac{[\text{Original value} - \text{Minimum value}]}{[\text{Maximum value} - \text{Minimum value}]}$$

where, $0 \leq N \leq 1$

Advantage: Neural Network lies in their ability to outperform nearly every other Machine Learning algorithms, but this goes along with some disadvantages that we will discuss and lay our focus on during this post. Like I already mentioned, to decide whether or not you should use Deep Learning depends mostly on the problem you are trying to solve with it. For example, in cancer detection, a high performance is crucial because the better the performance is the more people can be treated. But there are also Machine Learning problems where a traditional algorithm delivers a more than satisfying result.

Linear Regression - *Mansural Bhuiyan, Mohammad Al Hasan*
To establish baseline performance with a linear classifier, we used Linear Regression to model the price targets, Y, as a linear function of the data, X

$$f_{\mathbf{w}}(\mathbf{X}) = w_0 + w_1 x_1 + \ldots + w_m x_m$$
$$= w_0 + \sum_{j=1:m} w_j x_j$$

Advantage: A linear model can include more than one predictor as long as the predictors are additive. the best fit line is the line with minimum error from all the points, it has high efficiency but sometimes this high efficiency created.
Disadvantage: Linear Regression Is Limited to Linear Relationships. Linear Regression Only Looks at the Mean of the Dependent Variable. Linear Regression Is Sensitive to Outliers. Data Must Be Independent

Support Vector Regression - *C. Cortes and V. Vapnik*
We used the linear SVR and also the polynomial and Gaussian kernels for regression of target prices. The linear SVR estimates a function by maximizing the number of deviations from the actually obtained targets Yn within the normalized margin stripe, while keeping the function as flat as possible. In other word, the magnitude of the error does not matter as long as they are less than, and flatness in this case means minimize w. For a data set of N target prices with M features, there are feature vectors Xn ∈ RM where n = 1,...,N and the targets Yn corresponding to the price of real estate properties. The SVR algorithm is a convex minimization problem that finds the normal vector w ∈ RM of the linear function as follows.

$$\min_{\mathbf{w},\gamma} \left(\frac{1}{2}||\mathbf{w}||^2 + C \sum_{n=1}^{N} \gamma_n + \gamma_n^\star \right)$$

subject to the constraints for each n**:**

$$y_n - (\mathbf{w} * \mathbf{x}_n) \leq \epsilon + \gamma_n,$$
$$(\mathbf{w} * \mathbf{x}_n) - y_n \leq \epsilon + \gamma_n^\star,$$
$$\gamma_n, \gamma_n^\star \geq 0$$

Advantages: It has a regularization parameter, which makes the user think about avoiding over-fitting. Secondly it uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel. Thirdly an SVM is defined by a convex optimization problem (no local minima) for which there are efficient methods (e.g. SMO). Lastly, it is an approximation to a bound on the test error rate, and there is a substantial body of theory behind it which suggests it should be a good idea.

Disadvantages: Theory only really covers the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel. In a way the SVM moves the problem of over-fitting from optimizing the parameters to model selection. Sadly, kernel models can be quite sensitive to over-fitting the model selection criterion.

k-Nearest Neighbours (KNN) - *E. Fix and J. L. Hodges Jr*

k-Nearest-Neighbour (KNN) is a non-parametric instance-based learning method. In this case, training is not required. The first work on KNN was submitted by *Fix & Hodges* in 1951 for the United States Air-force. The algorithm begins by storing all the input feature vectors and outputs from our training set. For each unlabelled input feature vector, we find the k nearest neighbours from our training set. The notion of nearest uses Euclidean distance in the m-dimensional feature space. For two input vectors x and w, their distance is defined by:

$$d(\mathbf{x}, \mathbf{w}) = \sqrt{\sum_{i=1}^{m} (x_i - w_i)^2}$$

Advantages: The K-Nearest Neighbour (KNN) Classifier is a very simple classifier that works well on basic recognition problems.

Disadvantage: KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. To predict the label of a new instance the KNN algorithm will find the K closest neighbours to the new instance from the training data, the predicted class label will then be set as the most common label among the K closest neighbouring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data. Further, changing K can change the resulting predicted class label.

Random Forest Regression - *A. J. Smola and B. Sch¨olkopf*

The Random Forest Regression (RFR) is an ensemble algorithm that combines multiple Regression Trees (RTs). Each RT is trained using a random subset of the features, and the output is the average of the individual RTs. The sum of squared errors for a tree T is:

$$S = \sum_{c \in leaves(T)} \sum_{i \in C} (y_i - m_c)^2$$

$$\text{where } m_c = \frac{1}{n_c} \sum_{i \in C} y_i, \text{ the prediction for leaf } c.$$

Advantages: There is no need for feature normalization. Individual decision trees can be trained in parallel. Random forests are widely used. They reduce overfitting.

Disadvantages: They're not easily interpretable. They're not a state-of-the-art

5. Conclusion

We have defined several models with various features and various model complexities. There is a need to use a mix of these models a linear model gives a high bias (under fit) whereas a high model complexity-based model gives a high variance (overfit). Data Scientist tends to overfit their models which can be reduced by ridge regression and LASSO The study reveals that economic factors influence land price more than the social factors. The interaction of the selected factors (X) on land price (Y) is analyzed. It is found that four factors viz. GLV (84%), silver price per gram (92%), population (86%) and cost of crude oil (88%) have more positive effect on land price. The outcome of this study can be used in annual revision of guideline value of land which may add more revenue to the State Government while land transaction is made. This study will support the policy makers to relook the movement of the identified factors to have control on rise in the land price and stabilize it. Since there is a greater need for good long-term data analysis about land price, general land market behaviour and spatial development, the results produced in this research may be of great use for Government and non-Government agencies which involve in land administration.

References

[1] M. Praveena, V. Jaiganesh, International Journal of Computer Applications (0975 – 8887) Volume 169 – No.8, July 2017.
[2] Sampathkumar et al. / Procedia Computer Science 57 (2015)112 – 121.
[3] Kilpatrick, J.A Factors Influencing CBD Land Prices. Journal of Real Estate; 2000, 25: 28-29.
[4] Wilson, I.D., Paris, S.D, Ware, J.A., & Jenkins, D.H. Residential Property Price Time Series Forecasting With Neural Networks.Journal of Knowledge-Based Systems; 2002, 15: 335-341.
[5] Mark, A.S., & John, W.B. Estimating Price Paths for Residential Real Estate. Journal of Real Estate Research; 2003: 25, 277–300.
[6] Wang, J., & Tian, P. Real Estate Price Indices Forecast by Using Wavelet Neural Network, Computer Simulation, 2005:2.
[7] Zhangming, H. Research on Forecasting Real Estate Price Index Based on Neural Networks. Journal of the Graduates Sun Yat Sen University, 2006;27.
[8] Tinghao,. Real Estate Price Index Based on ARMA Model, Statistics and Decision; 2007, 7.
[9] David, E.D., & Paavo, M. Urban Development and Land Markets in Chennai, India. International Real Estate Review; 2008, 11: 142165.
[10] Steven, P., & Albert, B.F. Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research; 2009, 31: 147-164.
[11]Sampathkumar.V and Helen Santhi.M. Artificial Neural Network Modeling of Land Price at Sowcarpet in Chennai City, International Journal of Computer Science & Emerging Technologies; 2010, 1:44–49. Available at http:// download. excelingtech.co. uk/Journal/IJCSET%20V1%284%29.pdf
[12] Urmila, S. Module No. 3, Ports Logistics and Connectivity with Inland Container Depot, Institute of Rail Transport, Delhi; 2010, 61.
[13] Mansural Bhuiyan and Mohammad Al Hasan (2016) "Waiting to be Sold: Prediction of TimeDependent House Selling Probability" IEEE International Conference on Data Science and Advanced Analytics pp468-477
[14] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, The elements of statistical learning, 1. Springer, 2009, vol. 2
[15] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 0885-6125. DOI: 10.1023/A:1022627411411.
[16] V.N.Vapnik,TheNatureofStatisticalLearningTheory. New York, NY, USA: Springer-Verlag New York, Inc., 1995, ISBN: 0-387-94559-8.
[17] A. J. Smola and B. Sch¨olkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199–222, Aug. 2004, ISSN: 0960-3174. DOI: 10.1023/B:STCO.0000035301.49549.88. [Online]. Available: http://dx.doi.org/10.1023/B:STCO. 0000035301.49549.88.
[18] E. Fix and J. L. Hodges Jr, "Discriminatory analysisnonparametric discrimination: consistency properties," DTIC Document, Tech. Rep., 1951.

[19] IOP Conf. Series: Materials Science and Engineering 263 (2017) 042098 doi:10.1088/1757-899X/263/4/042098.
[20]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[21] S. C. Bourassa, E. Cantoni, and M. Hoesli, "Predicting house prices with spatial dependence: a comparison of alternative methods," Journal of Real Estate Research, vol. 32, no. 2, pp. 139–160, 2010. [Online]. Available: http://EconPapers.repec.org/RePEc:jre:issued:v:32:n:2: 2010:p:139-160.

