IMPROVED ACCURACY IN LAND PRICE PREDICTION USING RANDOM FOREST REGRESSION OVER NOVEL LASSO REGRESSION

K.S. Ugesh Kumar[1], Rashmita Khilar[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To enhance the accuracy in land price prediction using Random Forest Regression and Novel Lasso Regression.
Materials and Methods: This study contains 2 groups i.e Random Forest Regression and Novel Lasso Regression. Each group consists of a sample size of 10 and the study parameters include alpha value 0.05, beta value 0.2, and the power value 0.8. Their accuracies are compared with each other using different sample sizes also.
Results: The Random Forest is 88.39% more accurate than the Novel Lasso Regression of 75.44% in Land Price Prediction. The statistical significance difference (two-tailed) is 0.04 ($p<0.05$).
Conclusion: The Random forest model is significantly better than the Novel Lasso Regression in identifying Land Price Prediction. It can be also considered as a better option for the House Price Prediction.

Keywords: Random Forest, Novel Lasso Regression, Land Price Prediction, Accuracy, Machine learning , Samples.

[1]Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.
[2*]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu,India. Pincode: 602105.

1. Introduction

Land Price Prediction is the place where buyers and sellers can legally gamble the values of land to gain or lose. Housing prices are increasing rapidly, yet the numerous websites online where houses are sold or rented are less likely to be updated on a regular basis (Fei 2020). To predict the property price we described two different architectures, based on Neural Network and Decision Tree (Shinde and Gawande 2018). Similar applications of land price prediction are House Price Prediction, Property Price Prediction (Lai, Siu-fun, and Lai, n.d.), Real Estate Price Prediction (Zheng 2017).

In Land Price Prediction using Random Forest Regression related articles around 32 articles in Google Scholar, Sciencedirect and 21 in Scopus (Er 2018) The paper focuses on various prediction techniques used to forecast the property price using a neural network and decision tree (Lai, Siu-fun, and Lai, n.d.) This paper tells about both supervised and unsupervised learning. These two learnings combined to find the property value (Carr and Smith 1975). In this paper they are comparing linear regression, svm, k-nearest and multilayer and neural networks to predict the missing price in the real estate market and agent based simulation (Kundu 2019).In this study, supervised learning is combined with unsupervised learning to bridge this gap. To depict the uncertainty in property assessment, a method based on principle component analysis, a common unsupervised learning tool, was created and applied. (Lee 2021). In this paper they are creating machine learning algorithms with consideration of various factors in the real estate market in real time (Oecd and OECD 2018).

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy et al. 2022).The research gap in Land Price Prediction is the availability of real time data sets and the accuracy to be improved. The selection of the algorithm also plays a vital role in land price prediction, So, this research focuses on improved accuracy in land price prediction using Random Forest Regression over Novel Lasso Regression.

2. Materials and Methods

This work is carried out in the Data Analytics Lab, Department of Information Technology at Saveetha School of Engineering. The study consists of two sample groups i.e Random Forest Regression and Novel Lasso Regression. Each group consists of 10 samples with pre-test power of 0.18. The sample size kept the threshold at 0.05, G power of 80%, confidence interval at 95%, and enrolment ratio as 1.

Data Preparation

To perform Land Price Prediction the real time data sets used are house_data1. The input data sets for the proposed work is house_data1.csv collected from kaggle.com ("Kaggle: Your Machine Learning and Data Science Community"). The data sets consist of 12 attributes and 1461 instances. The attributes of house_data1 are depicted in Table 1.

The attributes Lot Area, MasUrarea, Bsmt Fsf are independent attributes TotalBsmt, 1stFloor, 2ndFloor, GrLivArea, GarageArea, WoodDeckSF, OpenPorchSF, Sale Price are dependent attribute .

Random Forest Regression

Random forest techniques, often known as regression forests, can be used to predict both classification and regression. The fundamental approach is to create a large number of decision trees based on a random selection of data and variables, and then use those trees to create a class of dependent variables. The key benefit of applying this technique to my dataset is that it can manage missing values and keep the correctness of the missing data. There is also a low risk of overfitting the model, and we may expect high dimensionality when employing it on a big level dataset. The outcome of regression trees will be continuous as per the following equation (1).

$$RFfi_i = \frac{\Sigma_j fi_{ij}}{\Sigma j \epsilon all\ features, k \epsilon all\ trees^{normfi_{jk}}} \quad (1)$$

Random forest destruction is the limitation of a call tree algorithm. It generates predictions without Requiring Many configurations in packages. It provides an associate degree effective means of handling missing knowledge. It will turn out an affordable prediction while not hyper-parameter standardization. Pseudocode and Accuracy Values for the regression model is mentioned in Table 2 and Table 4.

Novel Lasso Regression

Novel Lasso Regression may be a regularization technique. It's used over regression methodology for a lot of correct prediction.The model uses shrinkage wherever knowledge values square measure contracted towards a central pointasthemean. Novel Lasso Regression Is An L1 regularization technique.

$$\sum_{i=1}^{n} (y_i - \sum_j x_{ij}\beta_j) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

The Lasso procedure encourages a straightforward,sparse model with the description of estimation given as per the equation (2). The actual style of regression is similar temperament for modules showing high levels after you need to automatize bound components of a model choice like variable choice, parameter agent. The input fortheNovel Lasso Regression methodology is that the House_data1.csv and outputs is identification of the clusters to cluster customers of comparable characteristics.Novel Lasso Regression is a regression model that uses the L1 Regularization approach. Ridge Regression is used when the L2 Regularization Technique is used.We'll study a lot regarding these within the later sections. Pseudocode and Accuracy Values for the regression model is mentioned in Table 3 and Table 5 below respectively.

The minimum requirement to run the softwares used here are intel core I3 dual core cpu@3.2 GHz , 4GB RAM , 64 bit OS, 1TB Hard disk Space Personal Computer and Software specification includes Windows 8 , 10 , 11 , Python 3.8 , and MS-Office.

The land value is predicted by the comparative method. The current value of the land is determined by the price at which a comparable property in the area was recently sold. Where, Land = The expense of acquiring or purchasing land.

Statistical Package for the Social Sciences Version 23 software tool for statistical analysis, a software application was used. For accuracy, an independent sample T-test was used. The SPSS Software programme was also used to calculate standard deviation and standard mean errors. The group statistical values of proposed and existing algorithms are included in the significance values of proposed and existing algorithms. From the dataset, the dependent variables are bsmt sqrt feet and the independent variables are LotArea.

3. Results

The group statistical analysis on the two groups shows Random Forest has more mean accuracy than Novel Lasso Regression and the standard error mean is slightly less than Random Forest. TheRandom Forest algorithm scored an accuracy of 88.39% and Novel Lasso Regression has scored 74.6%.The graphical representation of Random Forest and Novel Lasso Regression is figured out below in Fig. 1. The accuracies are recorded by testing the algorithms with 10 different sample sizes and the average accuracy is calculated for each algorithm.

In SPSS, the datasets are prepared using 10 as sample size for Random Forest and Lasso Method. Group id is given as a grouping variable and Lot area is given as the testing variable. Group id is given as 1 for Random Forest and 2 for Lasso Method. Descriptive Statistics is applied for the dataset in SPSS and shown in Table 6, Group statistics is shown in Table 7, Two Independent Sample T-Tests in Table 8.

4. Discussion

From the results of this study, Random Forest is proved to be having better accuracy than the Novel Lasso Regression model. RF has an accuracy of 88.39% whereas LR has an accuracy of 74.6%. The group statistical analysis on the two groups shows that Random Forest has more mean accuracy than Novel Lasso Regression and the standard error mean including standard deviation mean is slightly less than Random Forest.

This paper tells about both supervised and unsupervised learning. These two learnings combined to find the property value (Kumar et al. 2020). In this paper they are comparing linear regression, svm, k-nearest and multilayer and neural networks to predict the missing price in the real estate market and agent based simulation (Zheng 2017) . In this study, supervised and unsupervised learning are used in this work. To reflect the uncertainty in property assessment, a method based on principle component analysis, a prominent tool for unsupervised learning, was devised and implemented (Gertel 1988) In this paper they are creating machine learning algorithms with consideration of various factors in the real estate market in real time (Kalliola, Kapočiūtė-Dzikienė, and Damaševičius 2021).

Land Price Prediction for House dataset was performed using Random Forest modulation based categorization which gave an accuracy of 88.39% (Hasan, Mahmudul Hasan, and Islam 2020). Attention based Land Price Prediction paper provided an accuracy of 74.6% (Ho et al. 2020).

The limitation in this model is that the accuracy of RF may get affected due to the inconsistent data and difficulty in getting the right datasets for analysis. Most of the data is simulated from nature which is far from reality effective data preprocessing techniques, and the combination of RF with other machine learning algorithms such as LR and RF may give better accurate results in the future.

5. Conclusion

In this research, Land Price Prediction is performed for the House_Data1 data set where it

was found using Random Forest Method. The quality of datasets formed with good sale price value and accuracy is improved.

Declarations

Conflicts of Interest

No conflicts of interest in this manuscript.

Author Contributions

Author UK was involved in data collection, data analysis, data extraction, manuscript writing. Author RK was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Vee Eee Technologies Solution Pvt. Ltd., Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102142.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." Bioresource Technology 351 (May): 126956.

Carr, Jack, and Lawrence B. Smith. 1975. "Public Land Banking and the Price of Land." Land Economics. https://doi.org/10.2307/3144949.

Er, Emrah. 2018. Applications of Machine Learning to Agricultural Land Values: Prediction and Causal Inference.

Fei, Yue. 2020. California Rental Price Prediction Using Machine Learning Algorithms.

Gertel, Karl. 1988. Farmland Prices: An Example of Economic Forecasts, Uses, and Limitations.

Hasan, H. M. Mahmudul, H. M. Mahmudul Hasan, and Md Adnanul Islam. 2020. "Emotion Recognition from Bengali Speech Using RNN Modulation-Based Categorization." 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). https://doi.org/10.1109/icssit48917.2020.9214196.

Ho, Ngoc-Huynh, Hyung-Jeong Yang, Soo-Hyung Kim, and Gueesang Lee. 2020. "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network." IEEE Access. https://doi.org/10.1109/access.2020.2984368.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." Environmental Research. https://doi.org/10.1016/j.envres.2022.113095.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." Materials Chemistry and Physics. https://doi.org/10.1016/j.matchemphys.2022.125900.

Kalliola, Jussi, Jurgita Kapočiūtė-Dzikienė, and Robertas Damaševičius. 2021. "Neural Network Hyperparameter Optimization for Prediction of Real Estate Prices in Helsinki." PeerJ. Computer Science 7 (April): e444.

Kumar, Govind, Department of Computer Science & Engineering, Amity University, Haryana, India, Priyanka Makkar, Dr Yojna Arora, et al. 2020. "Real Estate Price Prediction." International Journal of Innovative Research in Computer Science & Technology. https://doi.org/10.21276/ijircst.2020.8.6.1.

Kundu, Amitabh. 2019. "Land Price Changes in Lucknow City, 1970 - 1990." Urban Land Markets and Land Price Changes. https://doi.org/10.4324/9780429431531-6.

Lai, Siu-Fun Rita, Siu-fun, and Rita Lai. n.d. "Housing Price and Government Land Policies." https://doi.org/10.5353/th_b3125825.

Lee, Changro. 2021. "PREDICTING LAND PRICES AND MEASURING UNCERTAINTY BY COMBINING SUPERVISED AND UNSUPERVISED LEARNING." International Journal of Strategic Property Management.

https://doi.org/10.3846/ijspm.2021.14293.

Oecd, and OECD. 2018. "Average Price and Rent Price for Agricultural Land, 1990-2016." https://doi.org/10.1787/9789264085268-graph11-en.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." Journal of Energy Storage. https://doi.org/10.1016/j.est.2022.104422.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102102.

Shinde, Neelam, and Kiran Gawande. 2018. "Survey on Predicting Property Price." 2018 International Conference on Automation and Computational Engineering (ICACE). https://doi.org/10.1109/icace.2018.8687080.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022. "Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." Computational Intelligence and Neuroscience 2022 (February): 5906797.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" Fuel. https://doi.org/10.1016/j.fuel.2022.123494.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." International Journal of Biological Macromolecules 209 (Pt A): 951–62.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Zheng, Yujing. 2017. Spatial Analysis with Applications on Real Estate Market Price Prediction.

TABLES AND FIGURES

Table 1. Attributes of House_data1 Data Set Description

S.No	Attribute	Data Type	Value	Description
1	Id	Integer	Primary Key	Identity Number
2	Total Bsmt SF	Integer	0-1700	Total Bsmt Sqrt
3	Garage Area	Integer	0-900	Total Garage Area
4	Open Porch SF	Integer	0-250	Total Open Porch Sqrt
5	SalePrice	Integer	1-310000	Overall Sales Price of Land

Table 2. Pseudocode for Random Forest

// I : Input dataset records
1. Import the required packages.
2. Convert the Data Sets into numerical values after the extraction feature.
3. Assign the data to X_train, y_train, X_test and y_test variables.

4. Using train_test_split() function, pass the training and testing variables.
5. Give test_size and the random_state as parameters for splitting the data using SVM training.
6. Calculate the accuracy of the model.
OUTPUT **//Accuracy**

Table 3. Pseudocode for Novel Lasso Regression

// I : Input dataset records
1. Import the required packages.
2. Convert the Data Sets into numerical values after the extraction feature.
3. Assign the data to X_train, y_train, X_test and y_test variables.
4. Using train_test_split() function, pass the training and testing variables.
5. Give test_size and the random_state as parameters for splitting the data.
7. Compiling the model using metrics as accuracy.
7. Evaluate the output using X_test and y_test function
8. Get the accuracy of the model.
OUTPUT **//Accuracy**

Table 4. Accuracy of Land Price Prediction us Random Forest

Model Sample Size	**Accuracy**
Training Split- 71%, Test Split -29%	88.06
Training Split- 72%, Test Split -28%	88.10
Training Split- 73%, Test Split -27%	88.12
Training Split- 74%, Test Split -26%	88.13
Training Split- 75%, Test Split -25%	88.24
Training Split- 76%, Test Split -24%	88.22
Training Split- 77%, Test Split -23%	88.04
Training Split- 78%, Test Split -22%	88.16
Training Split- 79%, Test Split -21%	88.33
Training Split- 80%, Test Split -20%	88.39

Table 5. Accuracy of Land Price Prediction us Novel Lasso Regression

Model Sample Size	**Accuracy**
Training Split- 71%, Test Split -29%	71.11
Training Split- 72%, Test Split -28%	70.90
Training Split- 73%, Test Split -27%	75.34
Training Split- 74%, Test Split -26%	75.44
Training Split- 75%, Test Split -25%	75.36
Training Split- 76%, Test Split -24%	75.12
Training Split- 77%, Test Split -23%	75.15
Training Split- 78%, Test Split -22%	73.65
Training Split- 79%, Test Split -21%	73.74
Training Split- 80%, Test Split -20%	73.92

Table 6. Descriptive Statistic analysis, representing Random Forest and Novel Lasso Regression

Algorithm	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Group1	20	1.00	2.00	1.5000	.51299
Accuracy	20	70.40	75.44	73.2210	1.92684
Error	20	24.56	29.59	26.7750	1.92604
Valid N (listwise)	20	-	-	-	-

Table 7. Group Statistical analysis, representing Random Forest(mean accuracy 88.39%, standard deviation 1.71422,1.90860) and Novel Lasso Regression(mean accuracy 74.60%,standard deviation 1.71442,1.90583)

Algorithm	**N**	**Mean**	**Std. Deviation**	**Std.Error Mean**
Accuracy Random Forest Lasso	10 10	73.9730 72.4690	1.71422 1.90860	.54208 .60355
Error Random Forest Lasso	10 10	26.0220 27.5280	1.71442 1.90583	.54215 .60268

Table 8. Independent Sample Tests results with confidence interval as 95% and level of significance is 0.04.

Accuracy	**F**	**Sig.**	**t**	**df**	**Sig**	**Mean Difference**	**Std. Error Difference**	**95% Conf. Interval Lower**	**95% Conf. Interval Upper**
Accuracy Equal variances assumed	0.743	0.04	1.854	18	.080	1.50400	.81125	-.20038	3.20838
Equal variances not assumed			1.854	17.796	0.80	1.50400	.81125	-.20178	3.20978
Error Equal variances assumed	0.731	0.04	-1.858	18	0.80	-1.50600	.81064	-3.20910	.19710
Equal variances not assumed			-1.858	17.802	0.80	-1.50600	.81064	-3.21046	.19846

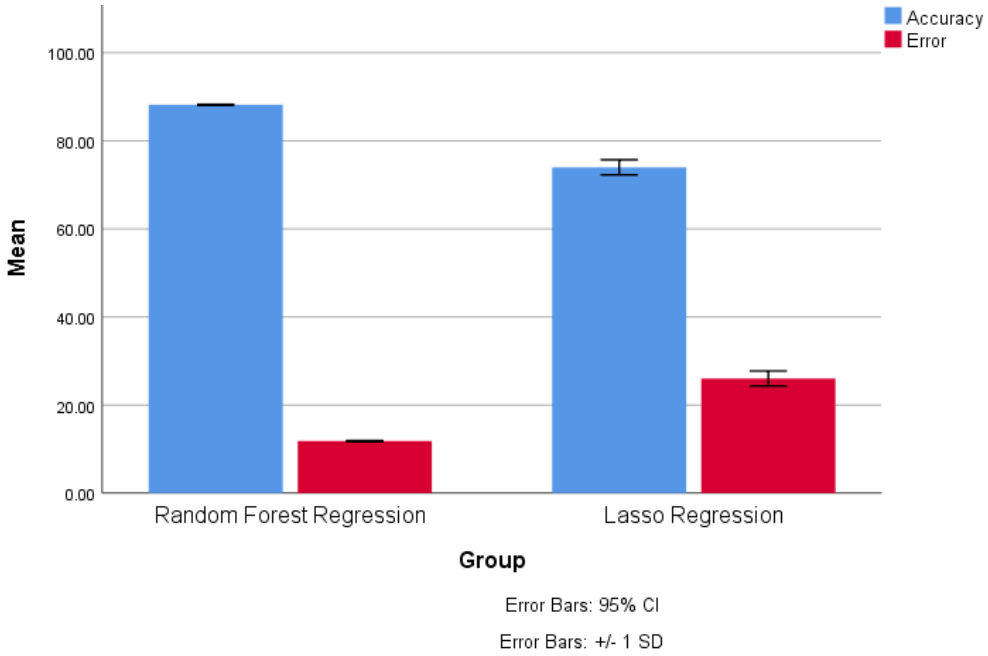

Fig. 1. Comparison of Random Forest Regression and Novel Lasso Regression in terms of accuracy. The mean accuracy of Random Forest Regression is greater than Novel Lasso Regression and the standard deviation is also slightly higher thanNovel Lasso Regression. X-axis: Random Forest Regression vs Novel Lasso Regression. Y-axis: Mean accuracy of detection + 1 SD