IMPROVED ACCURACY IN HOUSE PRICE PREDICTION USING LINEAR REGRESSION OVER RIDGE REGRESSION

B. Kirubakaran[1], Rashmita Khilar[2*]

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

Abstract

Aim: To improve the accuracy in House Price Prediction using Novel Linear Regression and Ridge Regression. **Materials and Methods:** This study contains 2 groups i.e Novel Linear Regression and Ridge 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 Gpower value 0.8.
Results: The Novel Linear Regression achieved accuracy (91.79) better than the Ridge Regression accuracy (91.34%) in House Price Prediction. The statistical significance difference (two-tailed) is 0.01 ($p<0.05$). **Conclusion:** The Novel Linear Regression model is significantly better than the Ridge Regression in House Price Prediction. It can be also considered as a better option for the House Price Prediction.

Keywords: Novel Linear Regression, Ridge Regression, House Price Prediction, Accuracy, Machine Learning, Sample Size.

[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

House Price Prediction can assist people in determining the selling price of a home and assisting customers in putting together the necessary funds at the appropriate moment to purchase the home. House price is usually predicted using house price index (HPI). It is commonly used to estimate the changes in house prices(Phan and The Danh Phan 2018). House price is correlated with factors such as location, area, population and also it requires other factors to predict the individual house price. House price forecasting is critical for determining, acquiring, and selling prices in a certain area (Madhuri, Anuradha, and Vani Pujitha 2019a) (Madhuri, Anuradha, and Vani Pujitha 2019b). (Mu, Wu, and Zhang 2014) proposed a model to predict the local house price using various machine learning techniques. Similar applications of House price prediction are Land Price Prediction, Property Price Prediction, Real Estate Price Prediction (Bala et al. 2020).

In House Price Prediction using Novel Linear Regression related articles around 87 in IEEE Digital Xplore and 92 in Science Direct (Phan and The Danh Phan 2018). There's been a lot of research into the housing market that contains data learning approaches (Phan and The Danh Phan 2018). Zhang's Paper focuses on the accuracy to evaluate house prices in all the states using Dynamic Model Average (DMA) and Dynamic Model Selection (DMS), two different forecasting techniques developed and motivated. The strategies take into account all of the K = 2m distinct version combinations in every occasion duration time t when there are m predictors available (Mu, Wu, and Zhang 2014). The techniques need utilizing K - modes methods to compute the possibility that model should be adopted for forecasting at time t. DMA considers the computed probabilities as model weights, while DMS adopts the model with the highest probability at time t (Bork and Møller 2015). Repeated loans transactions on same family properties whose mortgages were bought or securitized by Fannie Mae or Freddie Mac from January 1975 have been used to build this data (Truong et al. 2020).

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 House Price Prediction is the limited availability of real time data sets and the less accuracy in prediction. The selection of the algorithm also plays a vital role in house price prediction, So, this research focuses on improved accuracy in House Price Prediction using Novel Linear Regression over Ridge Regression.

2. Materials and Methods

This work is carried out in the Data Analytics Lab, Department of Information Technology at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The study consists of two sample groups i.e Novel Linear Regression and Ridge 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 house price prediction the real time data sets used are House Prices. The input data sets for the proposed work is USAHousing.csv collected from kaggle.com ("Kaggle: Your Machine Learning and Data Science Community" n.d.). The data sets consist of six attributes and 5000 instances. The attributes of House Prices are depicted in Table 1.

The attributes Average Area Income and Avg. Area Population attributes which are independent attributes which do not affect the results are removed from the .csv file.

Novel Linear Regression

Novel Linear Regression is a popular model, even with its simplistic representation. The representation is a linear equation that combines a chain of input values (x), with the solution being simply the forecast output for that set of input values (y). As a final result, both the enter (x) and output (y) values are numeric. Each input value or column is allocated one scale factor, known as a coefficient and represented by the capital Greek letter Beta in the linear equation (1). One more coefficient is brought, that allows the line an extra level of freedom (as an example, growing or reducing on a -dimensional plot) and is named as the intercept or bias coefficient.For instance, in a Novel Linear Regression situation with only one x and one y, the model would be:

$$y = B_0 + B_1 * X \qquad (1)$$

The line is called a plane or a hyper-plane in higher dimensions as there are several inputs (x). As a result, the representation consists of the equation's form as well as the coefficients' actual values(e.g. B0 and B1 in the above instance). The complexity of a regression version, along with Novel Linear Regression, is commonly discussed. The number of coefficients in the version is called this. When a coefficient equals 0, it successfully gets rid of the input variable's impact at the version and, as a result, the model's prediction (zero * x = zero). This is important to not forget while evaluating linearization techniques, which modify the gaining knowledge of algorithms to lessen the complexity of regression fashions via exerting stress at the absolute size of the coefficients or riding a few to 0. Pseudocode and Accuracy Values for the regression model is mentioned in Table 2 and Table 4.

Ridge Regression

Ridge regression is a model tuning approach which can be used to analyze data with multi - collinearity. L2 regularization is done using this method. If there is a problem with multi - collinearity, least-squares are unbiased, and variances are big, the projected values are far from the actual values. The cost function for ridge regression is given in the equation (2) as follows:

$$\mathrm{Min}(||\mathrm{Y} - \mathrm{X}(\theta)||^2 + \lambda||\theta||^2 \qquad (2)$$

The penalty term is lambda. The ridge function's alpha argument denotes the value supplied here. Varying the alpha values will be regulating the penalty term. The greater the alpha value, the greater the penalty, and hence the size of the coefficients is lowered. Pseudocode and Accuracy Values for the regression model is mentioned in Table 3 and Table 5.

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 house value is predicted by the comparative method. The current value of the house is obtained based on the size, age of the house and price of a house with similar amenities which is available in the same locality. House Price= Price Calculated using Sq.ft + year built + location + water facility + required amenities like no.of bedrooms, car parking to name a few.

Statistical Package for the Social Sciences Version 26 software was used for statistical evaluation. An independent sample T-test was carried out for accuracy. Standard deviation, standard mean errors has been additionally calculated using the SPSS Software tool. The significance values of proposed and existing algorithms contain group statistical values of proposed and existing algorithms. The independent variable is Area House Age, Avg. Number of rooms and Avg.Number of Bedrooms and accuracy and precision are dependent attributes.

3. Results

The group statistical analysis on the two groups shows Novel Linear Regression (group 1) has more mean accuracy than Ridge Regression (group 2) and the standard error mean is slightly less than Novel Linear Regression. The Novel Linear Regression scored an accuracy of 91.79% and Ridge Regression has scored 91.34%. The graphical comparison of the novel Linear Regression and Ridge Regression Model is figured 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 Novel Linear Regression and Ridge Regression. Group is given as a grouping variable and House Price is given as the testing variable. Group is given as 1 for Novel Linear Regression and 2 for Ridge Regression. 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, Novel Linear Regression is proved to be having better accuracy than the Ridge Regression model. Novel Linear Regression has an accuracy of 91.79% whereas Ridge Regression has an accuracy of 91.34%. The group statistical analysis on the two groups shows that Novel Linear Regression (group 1) has more mean accuracy than Ridge Regression (group 2) and the standard error mean including standard deviation mean is slightly less than Novel Linear Regression.

House Price Prediction using Machine learning is now becoming widely used as a methodology. Citizens who have employed machine learning algorithms to address problems based on their own industry data. Industry professionals have used machine learning to perform classification jobs and diagnose malfunctions. People in the field of business frequently used machine learning algorithms in financial research (Kaushal and Shankar, n.d.) The paper focus at the accuracy of evaluate housing prices in every 50 states using Dynamic Model Averaging (DMA) and Dynamic Model Selection (DMS), specific forecasting techniques developed and prompted by Raftery, Karny, and Ettler (2010) and Koop and Korobilis, correspondingly in the year 2012. The strategies take into account all of the K = 2m distinct model comparing time period t when there are m predictors available (Madhuri, Anuradha, and Vani Pujitha 2019a).(Madhuri, Anuradha, and Vani Pujitha 2019b).

The limitation of the proposed work is due to inconsistent data and difficulty in getting the right datasets for analysis. The future work can be concentrated on effective data preprocessing techniques and usage of ensemble machine learning algorithms can be focussed.

5. Conclusion

Based on the experimental results, the Novel Linear Regression has been proved to predict the house price more significantly than Ridge Regression. The quality of datasets formed with selling price value and accuracy is improved.

Declarations

Conflicts of Interest

No conflicts of interest in this manuscript.

Author Contributions

Author BK 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.

Bala, Ravula, Kunamneni Surya, Tadiparthi Chandravas, and Manikandan J. 2020. "A Machine Learning Based Advanced House Price Prediction Using Logistic Regression." International Journal of Computer Applications. https://doi.org/10.5120/ijca2020920303.

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.

Bork, Lasse, and Stig V. Møller. 2015. "Forecasting House Prices in the 50 States Using Dynamic Model Averaging and Dynamic Model Selection." International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2014.05.005.

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.

"Kaggle: Your Machine Learning and Data Science Community." n.d. Accessed March 25, 2021. https://www.kaggle.com/.

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.

Kaushal, Anirudh, and Achyut Shankar. n.d. "House Price Prediction Using Multiple Linear Regression." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3833734.

Madhuri, C. H. Raga, G. Anuradha, and M. Vani Pujitha. 2019a. "House Price Prediction Using Regression Techniques: A Comparative Study." 2019 International Conference on Smart Structures and Systems (ICSSS). https://doi.org/10.1109/icsss.2019.8882834.

———. 2019b. "House Price Prediction Using Regression Techniques: A Comparative Study." 2019 International Conference on Smart Structures and Systems (ICSSS). https://doi.org/10.1109/icsss.2019.8882834.

Mu, Jingyi, Fang Wu, and Aihua Zhang. 2014. "Housing Value Forecasting Based on Machine Learning Methods." Abstract and Applied Analysis. https://doi.org/10.1155/2014/648047.

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.

Phan, The Danh, and The Danh Phan. 2018. “Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia.” 2018 International Conference on Machine Learning and Data Engineering (iCMLDE). https://doi.org/10.1109/icmlde.2018.00017.

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.

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.

Truong, Quang, Minh Nguyen, Hy Dang, and Bo Mei. 2020. “Housing Price Prediction via Improved Machine Learning Techniques.” Procedia Computer Science. https://doi.org/10.1016/j.procs.2020.06.111.

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.

TABLES AND FIGURES

Table 1. USA Housing dataset collected from Kaggle Inc.

S.No	Attribute	Data Type	Value	Description
1.	Avg House Age	Integer	Primary Key	Age of the House
2	Avg. Number of Rooms	Integer	0-10	Number of Rooms in the House
3	Area Population	Integer	Percentage	Average Population in the Area
4	Price	Integer	More than 1 lakh in $	Price of the House

Table 2. Pseudocode for Novel Linear Regression

// I : Input housing dataset records
1. Import the packages whichever is required.
2. Convert all the variables in datasets into numerical values after the extraction feature.
3. Assign the training variables into data to X_train, y_train, X_test and y_test variables.
4. Using train_test_split() function, pass the values into both the train and test case variables.
5. Give the size of the test in variable test_size and the random_state as splitting parameter for splitting the data

```
and to use in the Linear Regression’s training model.
6. Compiling the model using metrics as accuracy
7. Calculate the accuracy of the model.
OUTPUT
//Accuracy
```

Table 3. Pseudocode for Ridge Regression

```
// I : Input housing dataset records
1. Import the packages whichever is required.
2. Convert all the variables in datasets into numerical values after the extraction feature.
3. Assign the training variables into data to X_train, y_train, X_test and y_test variables.
4. Using train_test_split() function, pass the values into both the train and test case variables.
5. Give the size of the test in variable test_size and the random_state as splitting parameter for splitting the data and to use in the Ridge Regression’s training model.
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 House Price Prediction using Novel Linear Regression

Model Sample Size	Accuracy
Training Split- 71%, Test Split -29%	91.79
Training Split- 72%, Test Split -28%	88.50
Training Split- 73%, Test Split -27%	90.04
Training Split- 74%, Test Split -26%	77.99
Training Split- 75%, Test Split -25%	85.79
Training Split- 76%, Test Split -24%	72.56
Training Split- 77%, Test Split -23%	91.55
Training Split- 78%, Test Split -22%	89.87
Training Split- 79%, Test Split -21%	82.25

Training Split- 80%, Test Split -20%	91.50

Table 5. Accuracy of Land Price Prediction using Ridge Regression

Model Sample Size	**Accuracy**
Training Split- 71%, Test Split -29%	91.74
Training Split- 72%, Test Split -28%	87.05
Training Split- 73%, Test Split -27%	89.06
Training Split- 74%, Test Split -26%	89.78
Training Split- 75%, Test Split -25%	82.45
Training Split- 76%, Test Split -24%	73.20
Training Split- 77%, Test Split -23%	82.90
Training Split- 78%, Test Split -22%	88.39
Training Split- 79%, Test Split -21%	77.80
Training Split- 80%, Test Split -20%	91.43

Table 6. Descriptive Statistic analysis, representing Novel Linear Regression and Ridge Regression

Algorithm	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Group1	20	1.00	2.00	1.5000	.51299
Accuracy	20	72.56	91.79	91.6420	6.22502
Error	20	8.21	27.44	9.3760	6.22474
Valid N (listwise)	20	-	-	-	-

Table 7. Group Statistic analysis, representing Novel Linear Regression (mean accuracy 91.65%, standard deviation 0.08600,0.09333) and Ridge Regression(mean accuracy 91.59%, standard deviation 0.08600,0.09333)

Algorithm	**N**	**Mean**	**Std. Deviation**	**Std.Error Mean**
Accuracy Novel Linear Regression Ridge Regression	10 10	91.6580 91.5900	0.08600 .09333	0.02719 .02951
Error Novel Linear Regression Ridge Regression	10 10	9.3420 9.4100	.08600 .09333	.02719 .02951

Table 8. Independent Sample Tests results with confidence interval as 95% and level of significance as 0.05 (Novel Linear Regression appears to perform significantly better than Ridge Regression with the value of p=0.05).

Levene's Test for Equality of Variances				**T-test for Equality of Means**					**95% Confidence Interval of the Difference**	
		F	**Sig.**	**t**	**df**	**Sig**	**Mean Difference**	**Std. Error Difference**	**Lower**	**Upper**
Accuracy	Equal variances assumed	.101	0.05	1.694	18	.080	.06800	.04013	-.01632	.15232
	Equal variances not assumed	-	-	1.694	17.881	.080	.06800	.04013	-.01636	.15236
Error	Equal variances assumed	.101	0.05	-1.694	18	.080	-.06800	.04013	-.15232	.01632
	Equal variances not assumed	-	-	-1.694	17.881	.108	-.06800	.04013	-.15236	.01636

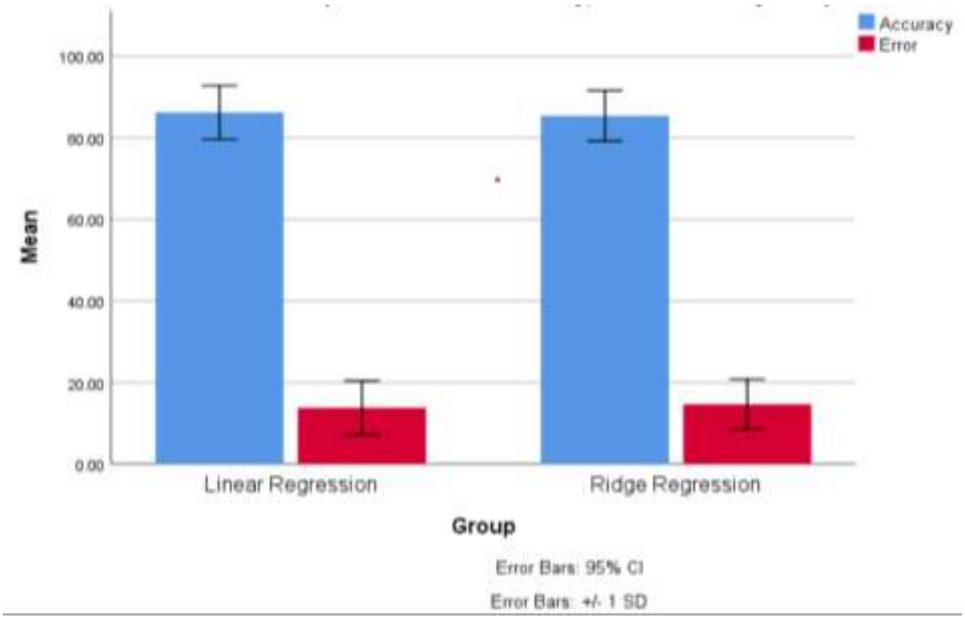

Fig. 1. Comparison of Novel Linear Regression and Ridge Regression with the mean term as accuracy. The mean accuracy of Novel Linear Regression is greater than Ridge Regression and the standard deviation is also slightly higher than Ridge Regression. X-axis: Novel Linear Regression vs Ridge Regression. Y-axis: Mean accuracy of detection + 1 SD.