Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **685:691***

House Prices Prediction Using Statistics with Machine Learning

Loai Nagib Alqubati[1*], Kiran Kumari Patil Loai Nagib[2]

[1,2]School of CSE, REVA University, India

**Corresponding author's E-mail: loai.nagib21@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 29 Nov 2023

Abstract

After the housing crisis in 2009 that affected the global economy and the bubble that burst, researchers began to focus on how to estimate house prices. In the United States, for instance, they adopted the hedonic price index (HPI) method in estimating house prices. After Ames house pricing dataset was released, which contains houses data from 2006 to 2010, and detailed features that help in studying the estimation of house prices. In this paper, we suggest that House prices are determined by many features such as area, utilities, house style, location, age, grade living area, number of bedrooms, garage, and so on. Statistical methods were applied with two models which are multiple and stepwise linear regression, also, two machine learning algorithms which are LASSO and XGBoost regression. The accuracy of prediction was evaluated by the root mean square error (RMSE). XGBoost with 25 features, 0.973 R2, and 0.027 RMSE is the Best model. LASSO has helped in feature selection for XGBoost.

Keywords: *House price prediction, SPSS, Feature Engineering, Machine Learning, XGBoost, and Lasso*

1. Introduction

One of the big financial decisions [14] that people make in their lives is to buy a house [4]. The development of civilization is the main reason for the increase in housing demand, resulting in prices increasing too [6]. This, in turn, results in people trying to keep an eye on house pricing and trying to figure out a more accurate way to predict the house pricing trends that take place at different times [5]. But how to use machine learning algorithms to predict housing prices? Not to mention, it is also a challenge to get accurate results.

The housing market is growing rapidly, and prices change a lot too depending on various factors, so house pricing is a real trend problem, therefore an important motivating factor for predicting house prices.

If you were looking for a home before the economic crisis in 2009, you could choose from the group of loans provided by banks to maintain your payments [4]. The traditional home price forecast is based on cost and sales price comparison, therefore there is a need to build a model to efficiently predict house prices, this model uses different features to describe each house, and every feature contributes to the price.

The house is determined by location, size, area, rooms, and other features [7]. In this paper, we are going to explain the methodologies that have been used. We have divided the work in this paper into 4 phases, which are Data description, exploratory data analysis (EDA) which is the approach of analyzing or initial investigation on datasets to test the hypothesis by SPSS (Statistical Package for the Social Sciences), machine learning algorithms such as LASSO (Least Absolute Shrinkage and Selection Operator), XGBoost (eXtreme Gradient Boosting) regression and feature engineering. "IBM SPSS Statistics is a powerful statistical software platform. It delivers a robust set of features that helps your organization extract actionable insights from its data." [1]. We used Statistical methods in SPSS for preprocessing, preparing datasets, handling missing values, Outliers, and feature selection. We have built 2 models in SPSS which are Multiple linear regression and stepwise multiple regression. Stepwise based on the probability of F (P-value), SPSS starts entering the independent variables with the smallest P-value 0.05 then in the next step again the variable from the list not in the equation yet with the smallest P-value and so on [3].

The features with the highest P-value will be removed from the model [3]. Many regression algorithms use for building a model and predict house prices [7], we found LASSO with shrinkage and XGBoost with ensemble learning give good results.

This paper is organized as related work in section II, methodology in section III, results in section IV, and conclusion in section V.

Related Work

Lasso regression algorithm is a regression method that performs feature selection, regularization and adds L1 penalty terms [7],[9],[10],[11]. After selecting the alpha parameter, the LASSO algorithm can list coefficients zero features [7] or the coefficients estimates of the explanatory variable towards zero [8]. In [11] got the best RMSE 0.11139 with LASSO regression. Also, In [12] got 0.11449 as RMSE result. XGBoost is a machine learning algorithm free tree boosting [12]. In. [11] trained 100 trees with a maximum depth of 20.

In [12] investigate that XGBoost runs more than 10 times faster than other methods with result RMSE result of 0.16118.

The good results of the LASSO algorithm when combined with XGBoost as a hybrid regression model [7],[12].

For missing values imputation, A few documents explore that statistical methods are easiest and fastest, according to the paper by [7] and [11] they filled categorical missing values with mode and numerical values with the median. On the other hand when the missing values are huge more than 55% removing a feature will be a good solution [6],[12].

In [10] and in [11] investigate the heteroscedasticity of numerical features and use Log transformation to deal with them. Log transformation to approximate normal distribution [7].

For expansion features, the majority of datasets contain various data types including numerical features, either continuous or discrete, and categorical features, either nominal or ordinal, according to [10] The most influential component with the highest frequency is highlighted by converting nominal characteristics with severely uneven distribution of values to binary variables. The expansion is also more flexible for numeric features. In [7] investigate that the more features, the better score for which they added 400 more new features.

LASSO regression is a good way to choose the best features and drop the useless features, while XGBoost is characterized by high speed and accuracy in the results. After looking at previous research papers related to housing price prediction, we find that combine Lasso and XGBoost would give the best scores.

2. Materials and Methods

Data Description

The Ames Housing dataset from Kaggle [2] consists of records ranging from 2006 to 2010. This dataset contains 81 features, including the dependent feature “SalePrice” and 2919 records, among them, 20 features are Continuous, 14 discrete, 24 nominal, and 23 ordinal. This dataset is popular data in house pricing prediction from Ames city, Iowa, USA. Besides that we have combined the school’s distances with Longitude and Latitude for each home there are 8 schools in Ames which are 5 elementary, one middle, and one high school, so the total of features in the dataset will be 92 features.

Exploratory Data Analysis by SPSS

Ordinal data is denoted using sequential numbers (Orders of numbers). ‘BsmtQual’ features have 5 values (Ex, Gd, TA, Fa, Po, NA) so we change to as shown in table 1:

TABLE 1. Ordinal Values

String values	Order values
Ex	5
Gd	4
TA	3
Fa	2
Po	1
NA	0

For nominal data, one hot-encoding or dummy code is used. Which replaces each instance in the features with a number that helps in identification as shown in table 2:

TABLE 2. After Applying One-Hot Encode.

INDEX	LOTCONFIG_ INSIDE	LOTCONFIG_FR2	LOTCONFIG_CORNER
0	1	0	0
1	0	1	0
2	1	0	0
3	0	0	1

In the dataset, there are 824 missing values. 74 missing values were filled manually from the Ames city assessor's official website. The numerical features contain 677 missing values. The categorical features contain 73 missing values. Table 3 has been shown the highest missing values.

TABLE 3. MISSING VALUES.

FEATURES	TYPE	MISSING VALUES
Lot Frontage	Continuous	486
Mas Vnr Type	Nominal	23
Mas Vnr Area	Continuous	23
Exter Cond	Ordinal	3
Bsmt Qual	Ordinal	3
Functional	Ordinal	4
Garage Yr Blt	Discrete	158
Garage Fin	Ordinal	2
Garage QC	Ordinal	4
Garage Cond	Ordinal	4
Pool QC	Ordinal	13
Misc Feature	Nominal	6

Missing values must be filled in to make the model work probably. Numerical features were filled with median value and all categorical features were filled with mode value [7],[11], [15-20].

Correlations are used to express the strength of the relationship between two or more variables. Significance level or P-value tells how well the data reject the null hypothesis. After recognizing correlations, features that have a significance level (sig) more than 0.05 were removed.

(1)

After correlation we found 41 features have P-values greater than 0.05, hence are considered not important, as a result, such features were removed.

For outliers, we used Mahalanobis $t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$ distance for detecting outliers in the dataset. Mahalanobis distance is usually used for multivariate outliers as this distance considers the shape of features.

We investigate the heteroscedasticity of numerical features using Shapiro-Wilk [8] and Kolmogorov-Sminov normality test and we used Log transformation to deal with heteroscedasticity [7],[10],[11]. Log transformation has been applied for ‘SalePrice’, ‘LotArea’ and ‘LotFrontage’ etc. as shown in fig. 1

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SalePrice	.123	2919	<.001	.875	2919	<.001

a. Lilliefors Significance Correction

FIG. 1. Normality distribution test

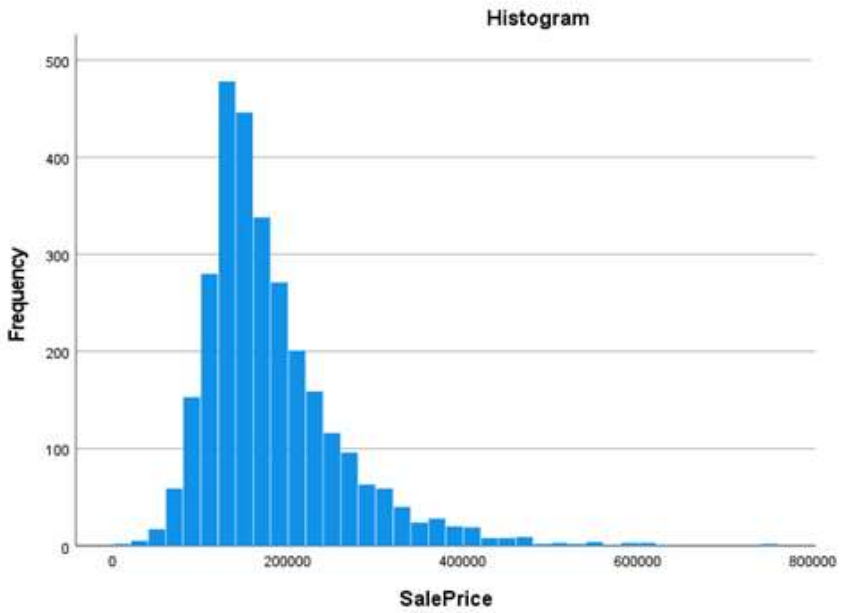

FIG. 2. Sales Price destiribution

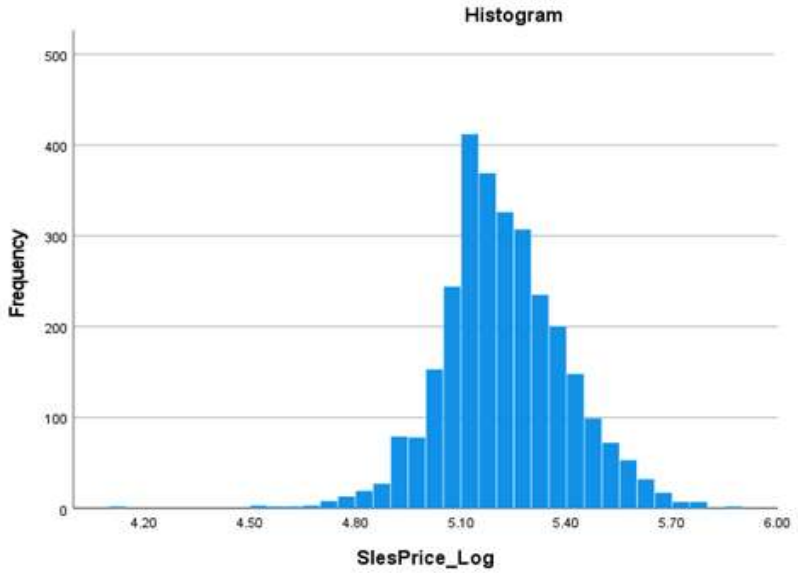

FIG. 3. Log (SalePrice)

Multiple linear regression is applied using 111 features.

Model Summary[b]

Model	R	R Square	Adjusted R Square
1	.965[a]	.931	.928

FIG. 4. Multiple linear regression results

- for stepwise multiple linear regression, 45 features were included and R2 is .928.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- RMSE is the standard deviation of the error (residual error). It's always positive, and smaller RMSE is better.
- Models were evaluated using R2 and Root Mean Square Error RMSE, as shown below.

$$R^2 = 1 - \frac{\sum_i (yi - \hat{y}i)^2}{\sum_i (yi - \bar{y})^2}$$

Machine Learning Algorithms

LASSO regression implements the L1 regularization technique, also LASSO helps to build the generalized model with low bias and low variance and adds "absolute value of magnitude" of coefficient as penalty term to the loss function or cost function.

$$\sum_{i=1}^{n} (y - \hat{y})^2 + \lambda |slope|$$

Suppose if the model is

$$\lambda |m_1 + m_2 + m_3 + \cdots + m_n|$$

The magnitude of the slope is nothing but

$$y = \mathrm{m}_{1x1} + \mathrm{m}_{2x2} + \mathrm{m}_{3x3} + \cdots + \mathrm{m}_{nxn} + c$$

If the slope values are very less those features will be removed, so LASSO helps us to select features. We selected parameter alpha; fig.5 shown

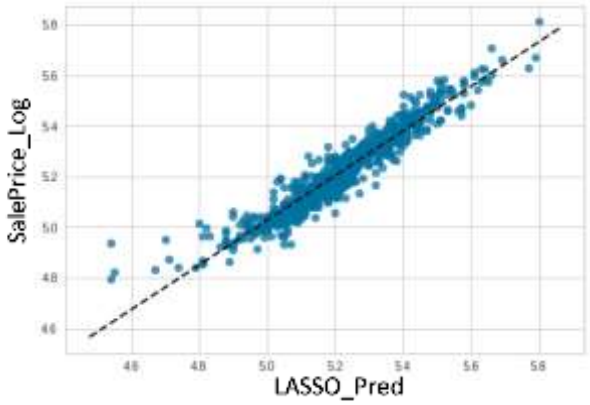

FIG. 5. LASSO prediction

Another algorithm is XGBoost which is an ensemble technique using boosting technique, it is a very popular algorithm. We performed hyperparameters tuning include n_estimators, max_depth, learning_rate, subsample min_child_weight, gamma, colsample_bytree.

The idea behind XGBoost is that the creation of multiple decision trees and similarity weights for each tree is calculated as follows:

$$similarity\ weight = \frac{\sum (Res)^2}{No\ of\ Res + \lambda}$$

After that Gain is calculated

$$Gain = Left\ similarity + Right\ similarity - Root\ similarity$$

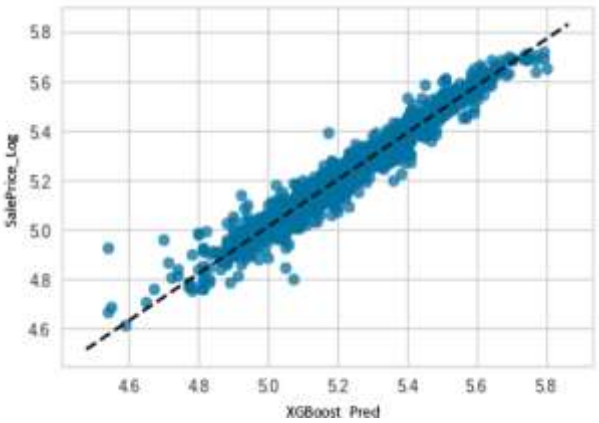

FIG. 6. XGBoost prediction

Feature Selection

The main idea of this study is based on the optimal selection of the features that have an impact on the house sales price. In multiple linear regression, the number of features is 111, stepwise linear regression 45. For LASSO regression number of features has been reduced to 25 after that XGBoost model is built with LASSO features to enhance the result.

3. Results and Discussion

TABLE 4. Results of Algorithms

Machine Learning Algorithms	No of features	R2	RMSE
Multiple linear regression	111	0.928	0.0451
Stepwise linear regression	45	0.907	0.0446
LASSO	25	0.910	0.0505
XGBoost	25	0.973	0.0271

Many models were created for the best selection of features as we see table above, number of features in Multiple linear regression 111 but after stepwise method in SPSS the number has reduced to 45. 10-fold-cross validation is used to test the evaluation metrics and choose optimal alpha and, and we selected alpha=0.005 and model has chosen 25 features and dropped other. In XGBoost 25 features are selected (LASSO features) with hyperparameters tuning include Colsample_bytree=0.2, while learning_rate=0.05 and subsample=0.2, also the score increased from 0.910 to 0.973. After investigation, that best model is XGBoost with 25 features, 0.973 R2 and 0.027 RMSE.

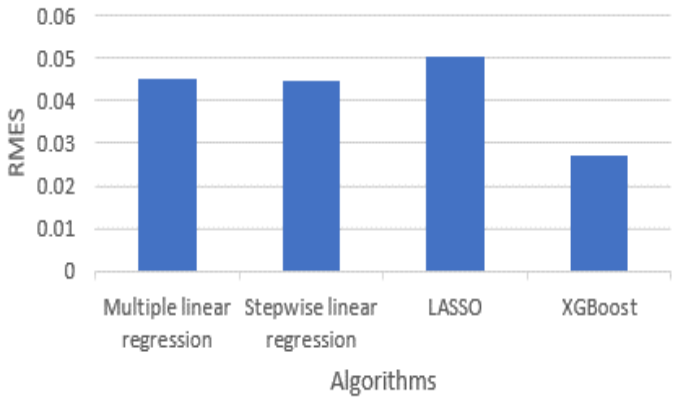

FIG. 7. Comparison of models

4. Conclusion

In this paper, the solution for the house pricing prediction problem was introduced, in which we used statistical methods and machine learning algorithms. This project aims to provide a model through which people or real estate agents can forecast house prices. We were able to process the missing data, make efforts in feature engineering to choose the best features. What's more, compare the results that we have got from models. LASSO has helped in feature selection for XGBoost.

As future work on this paper, we could try to generalize this project to implement in other areas, improving the model by adding other features that impact house pricing such as Deposit rate, crime, and unemployment rate. A more knowledgeable prediction about how they would affect the different machine learning models such as Support Vector Regression can be used to run the experiment and

the Polynomial Regression model, also, using XGBoost with Ridge regression, combine LASSO and Ridge regression (elastic net regression).

Acknowledgment

I would like to express my sincere gratitude to my guide Prof. Kiran Kumari Patil provided me with the invaluable opportunity to work on this fantastic project on the topic of house pricing prediction, which also assisted me in conducting extensive research and learning many new concepts.

References:

[1] https://www.ibm.com/products/spss-statistics
[2] https://www.kaggle.com/c/ames-housing-data
[3] https://www.ibm.com/support/knowledgecenter/SSLVMB_sub/statistics_mainhelpddita/spss/base/linear_regression_methods.html
[4] https://www.washingtonpost.com/news/business/wp/2018/10/04/feature/10-years-later-how-the-housing-market-has-changed-since-the-crash/
[5] Liu, Z., Cao, J., Xie, R., Yang, J., & Wang, Q. (2020). Modeling Submarket Effect for Real Estate Hedonic Valuation: A Probabilistic Approach. IEEE Transactions on Knowledge and Data Engineering, 1–1. https://doi.org/10.1109/tkde.2020.3010548
[6] Zhongyun, J. (2019). Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework. In 2019. IEEE 6th International Conference on Systems and Informatics (ICSAI).
[7] Phan, T. D. (2019). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. Proceedings – IEEE International Conference on Machine Learning and Data Engineering, ICMLDE 2018, 8–13. https://doi.org/10.1109/iCMLDE.2018.00017
[8] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, R. S. M. G. (2017). A Hybrid Regression Technique for House Prices Prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).
[9] Oladunni, T., Sharma, S., & Tiwang, R. (2017). A spatio-temporal hedonic house regression model. Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2017-Decem, 607–612. https://doi.org/10.1109/ICMLA.2017.00-94
[10] Masrom, S., Mohd, T., Jamil, N. S., Rahman, A. S. A., & Baharun, N. (2019). Automated machine learning based on genetic programming: A case study on a real house pricing dataset. Proceedings - IEEE 1st International Conference on Artificial Intelligence and Data Sciences, AiDAS 2019, 48–52. https://doi.org/10.1109/AiDAS47888.2019.8970916
[11] Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. ACM International Conference Proceeding Series, 6–10. https://doi.org/10.1145/3195106.3195133
[12] Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I., & Vasilevna, P. I. (2018). Predicting sales prices of the houses using regression methods of machine learning. RPC IEEE - Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications, 1–5. https://doi.org/10.1109/RPC.2018.8482191
[13] Truong, Q., Nguyen, M., Dang, H., & Mei, B. Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174(2019), ELSEVIER 433–442. https://doi.org/10.1016/j.procs.2020.06.111.
[14] P. Mohan and K.K. Patil, "Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application", International Journal of Intelligent Engineering and Systems, Vol.11, No.4, pp.167-176, 2018.
[15] Sudhan Murugan Bhagavathi, Anitha Thavasimuthu, Aruna Murugesan, Charlyn Pushpa Latha George Rajendran, A Vijay, Raja Laxmi, Rajendran Thavasimuthu, Weather forecasting and prediction using hybrid C5.0 machine learning algorithm International Journal of Communication Systems, Vol. 34, Issue. 10, Pp. e4805, 2021.
[16] PM Surendra, S Manimurugan, A New Modified Recurrent Extreme Learning with PSO Machine Based on Feature Fusion with CNN Deep Features for Breast Cancer Detection, Journal of Computational Science and Intelligent Technologies, Vol. 1, Issue. 3, Pp. 15-21, 2020.
[17] PK Sadineni, Comparative Study on Query Processing and Indexing Techniques in Big Data, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 933-939, 2020.
[18] AH Omar Baabood, Prajoona Valsalan, Tariq Ahmed Barham Baomar, IoT Based Health Monitoring System, Journal of Critical Reviews , Vol. 7, Issue. 4, pp. 739-743, 2020.
[19] Sajay KR, Suvanam Sasidhar Babu, Vijayalakshmi Yellepeddi, Enhancing The Security Of Cloud Data Using Hybrid Encryption Algorithm, Journal of Ambient Intelligence and Humanized Computing, 2019. https://doi.org/10.1007/s12652-019-01403-1
[20] Bindhia K Francis, Suvanam Sasidhar Babu, Predicting academic performance of students using a hybrid data mining approach, Journal of Medical Systems, 43:162, 2019. https://doi.org/10.1007/s10916-019-1295-4