Retraction

Retraction: Parameterized Comparison of Regularized Regression Models to Develop Models for Real Estate (IOP Conf. Ser.: Mater. Sci. Eng. 1099 012016)

Published 05 July 2022

This article has been retracted by IOP Publishing following an allegation of overlap with another unpublished work by different authors. Upon investigation, IOP Publishing finds the article to be heavily plagiarised from multiple sources, making the work nonsensical. IOP Publishing is concerned this was not detected by the conference and believes the article may not have been legitimately peer reviewed. The authors of the article and conference organisers have all been unresponsive to our enquiries.

IOP Publishing has investigated in line with the COPE guidelines, and agree that this article should be retracted.

IOP Publishing Limited have not received a response from the authors regarding this retraction. The authors are encouraged to contact IOP Publishing Limited if they wish to contest this retraction.

Retraction published: 05 July 2022
Parameterized Comparison of Regularized Regression Models to Develop Models for Real Estate

Ankur Chaturvedi¹, Akshi Gupta² and Vikram Rajpoot¹

¹Assistant Professor, GLA University, Mathura, India, 303007
²Scholar, GLA University, Mathura, India, 303007

E-mail: ankur.chaturvedi@gla.ac.in

Abstract. This research paper has been made on the data of life expectancy. Data carries two sort of regression tasks in it; one is continuous feature (Life expectancy) while another one is discrete feature (Status). Life expectancy depends on many a thing such as alcohol consumption, polio, infant deaths, etc. Generally, in data there exists two models separately, but research has been made to implement both at once. Research goes in a manner that it also involves the comparison or models accuracy among linear, ridge, and lasso. Visualization, normalization, data cleaning, feature reduction, etc, is also performed so as to increase the accuracy. One always looks for less time to complete task with less workout. Ultimately, research successfully implemented both linear regression and logistic regression both at once with optimized model. It is also stating the importance of the ridge and lasso algorithms for optimization.

1. Introduction

In this entire growing population, there are several existing countries in the world having varied development criteria. Some countries, we know, are already developed with some advancement in technologies as well as in the hospitality services whereas, on the other hand, some countries are still developing [1] and growing in several fields with the growing world, they are also keeping pace. Countries’ development affects the infrastructure of Life expectancy based on some other factors as well. Life expectancy is, somewhat, the probability of life being expected after considering some factors that affect it. In our dataset, features like adult mortality, alcohol, diseases, expenditure, population, and many other are affecting life expectancy [2]. The following dataset is approved by World Health Organization (WHO). Predicting life expectancy comes under the category of predicting continuous values while prediction of status comes under the category of discrete values (categorical division) [3].

2. Literature Review

Sensitive factors in Life expectancy dataset affects the Life expectancy of a population, meanwhile, it gives different life expectancy report for different countries as factors may vary with the regions. Previous research possesses the theory to correlate the factors with the life expectancy of population predicted by different counties, focusing on the fact of perceiving the long-standing disease(s) affecting the life span of population, using Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) algorithms [4][5][6] also been classified on the some of the advanced algorithms that includes
Random Forest Classifier, and Decision Tree. The research compared two regression algorithms, SLR and MLR, and highlighted that the result proposed by Multiple Linear Regression algorithm over the prediction analysis (prediction of Life expectancy) is much appropriate than Simple Linear Regression. In contrast, our research model predicted two labels (Life expectancy and Status) using Lasso, Ridge, and Linear Regression. The basic approach is to utilize the algorithms to gain more outcomes.

3. Theoretical Framework

Some of the Machine Learning algorithms taken in use in this research:

3.1 Linear Regression

It is a Machine Learning algorithm which comes under supervised learning. It is that process of predicting a continuous value, used to find the linearity between one or more than one predictor [3] and target.

3.1.1 Simple Linear Regression

Predicting a response using a single feature: It is a method to predict the dependent variable (Y) based on the values of independent variables/features (X). Let’s assume that the two variables are linearly related. Hence, try to find the linear function through $y=mx+c$ that predicts the output value(y) that is probably a function of the feature or independent variable(x) with less error.

How to find the best fit/suitable line: In this simple linear regression model, trying to minimize the errors in prediction by finding the “best-fitted line”, the line from the errors would be minimal. Consequently, the goal is clear that the model aims at decreasing the difference between y_actual and y_predicted [4].

3.1.2 Multiple Linear Regression

Multiple Linear Regression is an algorithm that is used to model the relationship among more than one feature and responses the output by fitting an equation of Linear Regression [5] to observed data. There are some steps to perform the Multiple Linear Regression which is similar to Simple Linear Regression. The only difference between them is the way of evaluation. It proposes to find out the dependency of one feature over others and also let you know which factor is highly impacting the predicted output [7][8].

![Figure 1. Fitting n-Dimensional Data using Multiple Linear Regression](image-url)
Linear Regression is based on the rule of Ordinary Least Square (OLS) that states the sum of the squared residuals which is to be minimised:

$$
\min_w \|Xw - y\|^2
$$

(1)

Another two regression algorithms that are used in this research are Ridge & Lasso algorithms.

- **Ridge Regression**: It works similar to the Linear Regression means OLS. But additionally, it has advancement as the OLS + λ times \((slope)^2\).

- **Lasso Regression**: Another supervised alternate of Linear Regression. Additionally, it has advancement as the OLS + λ times \(|slope|\).

### 3.2 Logistic Regression

It is a machine learning algorithm that is used for classification problems. It follows a sigmoid path due to its function which can be written as

$$
Activation\ Function = \frac{1}{1+e^{-hypothesis}}
$$

(2)

What model going to do here is, creating a model using a logistic regression algorithm that just classifies the data between the event i.e., it is happening or not. Hence, we define a threshold value that predicts the plot under 0 and 1 as 0.5 being threshold value. It is based on the concept of probability and does the predictive analysis.

Logistic Regression is a classification algorithm that is used for problems having different classes. In this type of algorithm, we have to predict the output in discrete value (binary outcome either 0 or 1). For instance, we can predict whether a person can vote or not in the upcoming elections.

Logistic Regression is quite similar to the Linear Regression but the only difference between both is the nature of the desired output. The Linear Regression predicts the continuous value whereas the Logistic value predicts the discrete value. Therefore, Logistic Regression can also be known as Generalized Linear Regression.

**How does it work?**

Logistic Regression simply measures the relationship between the labels (dependent variable) which we have to predict and the features (independent variables), by probably estimating the outcomes using the underlying logistic function.

**Sigmoid Function**

The Sigmoid Function is an activation function that is nothing but an S-shaped curve which takes a real-valued number and then map it after applying a particular activation function on it. The sigmoid function maps the real values in between the range of 0 and 1 but never places the values exactly on those points.

### 4. Experimental Procedure

#### 4.1 Loading Data

Machine learning models in this research use House Price data from the Kaggle House Price Prediction Competition-House Prices: Advanced Regression Techniques. The datasets are in csv files. Train data and Test data both the datasets are loaded from the csv files.
4.2 Understanding Data

The training dataset contains 81 columns and 1460 rows that may affect estate prices. The test dataset contains 80 columns and 1459 rows. DataFrame.info function gives the concise summary of the data frame including list of all columns with their respective data types and the number of non-null values in each column. Sum of all the null values in each column helps to understand the need of each column in model. Dataframe.describe is used to view some of the basic statistical details like percentile, mean, standard deviation etc. of a data frame or a series of numeric values.

4.3 Data Visualization and Feature Selection

In this research, for the Data visualization Seaborn and Matplotlib libraries of Python are used. Representation of null values in train data and test data gives the information about null values in respective columns.

The skew nature of Sale Price column helps to understand about the asymmetry of the probability distribution of a real valued random variable around its mean, represented by Figure 2.

![Figure 2. Skew Nature of Sale Price Column](image)

The log of Sale Price column gives the proper normally distributed graph as shown in Figure 3.

![Figure 3. Normal Distribution after Taking Log](image)
The relationship between Quality and Sale Price is shown by below in Figure 4 graph and it is clear from graph that as the Quality increases the Sale Price also increases and it is very obvious in real time also.

**Figure 4.** Over Quality vs Sale Price

Figure 5 shows the Regression plot of Garbage Area with Sale Price shows that how the sales price of house gets effected in garbage area.

**Figure 5.** Garbage Area vs Sale Price

The Sale Price also vary with the Sale Conditions, conditions like Abnormal, Family, Partial, Normal, Adjacent Land etc. As these conditions get changes sale price get increased or decreased accordingly, this can be understood by Figure 6 below.
4.4 Building Models and Calculating R2-score, MSE, MAE

In this research, we had chosen regression machine learning as the prices of house are in continuous form. Three different models are made by using linear regression algorithm and R2-score, MSE, MAE are calculated. The implementation of algorithm is done with the help of scikit-learn library.

R2-score is defined as the proportion of variance in the responding variable that is predictable from the predictor variable or variables. The r2 score value varies in between 0 and 100%.

Mean Square Error [MSE] is positive value and closer to zero. MSE value must be lower value for better accuracy. This is calculated as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i')^2$$

Mean Absolute Error [MAE] is the sum of absolute difference between predicted and actual values. It mainly considers the direction, i.e., positive or negative value.

Formula for calculating mean absolute Error or MAE is given by:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - y_i'|$$

Finally, data[df] is made with no null values in it by dropping all the features having null values. Now there are 38 features present in dataset. Dataset is divided in the ratio of 80% as training data and 20% as testing data.

4.4.1 Building a Model with all the Features

Linear Regression model is made by taking all the features of dataset. X_train and y_train are fit over the model and then future values of house is predicted. Hence, R2-score, MSE, MAE are calculated. The Figure 7 below represents the relationship between Predicted price and Actual price.
4.4.2 Building a model by Lasso Regularization

Lasso regularization or L1 regularization/penalty is an algorithm which is used for estimating the generalized linear models. Lasso basically contains a penalty term that limits the size of the estimated coefficients. Lasso is also termed as shrinkage estimator because it generates coefficient estimates that are biased to be small. Also, L1 regularization can be served as built-in feature selection method. Here, we use this regularization and generate a model to compare R2-score, MSE, MAE.

4.4.3 Building a model using Ridge Regularization

Ridge regularization or L2 regularization/penalty is used for squaring the coefficients. Whereas, Lasso regularization is mostly used to take the absolute value of the coefficients. Therefore, ridge regularization puts the limits on the coefficients.

![Figure 7. Predicted Price vs Actual Price](image)

![Figure 8. Heatmap of Correlation Coefficient Values](image)
The penalty term which basically termed as lambda, regularizes the coefficients in such a way that the optimization function is penalized if the coefficients take up large values. So, ridge regularization shrinks the coefficients and it also helps to decrease the complexity of model and multi-collinearity. We develop the model by applying ridge regularization and compare different values of R2-score, MSE, MAE. Figure 8 perfectly represent that What is correlation between different coefficient values.

5. Results

Basically, there are three most common metrics that are used to measure the accuracy for the continuous variables i.e., R2-score, Mean Squared Error [MSE], Mean Absolute Error [MAE]. Table given below represent the MSE, MAE and R2-score calculated in all three models implemented above. This gives an idea that which model gives more accuracy and which will provide less accuracy. As, from the Table 1, we can observe that all the features are taken then the model is having highest accuracy among all three cases.

| MODEL                              | R2-SCORE | MSE          | MAE          |
|------------------------------------|----------|--------------|--------------|
| Linear Regression Model taking all features | 0.9532050323789 | 0.0564366108907 | 0.0079869678705 |
| Linear Regression Model using Lasso Regularization | 0.8564344988807 | 0.1141284108581 | 0.0232105689095 |
| Linear Regression Model using Ridge Regularization | 0.9120827181618 | 0.0923417433734 | 0.0150057269188 |

This shows that as number of features increases the accuracy also increased accordingly. Ridge regularization also show good accuracy.

6. Conclusion

In this paper, we measure accuracy by using different criteria over linear regression machine learning algorithm to know which will give better result for this specific dataset. We have determined the performance of every model and then we can analyze from the above Table easily. By running these three models, Linear regression model with all the features taken has shown the highest accuracy of 95.32%. Also, Linear regression model using Lasso Regularization and Linear regression model using Ridge Regularization has shown 85.64% and 91.21% accuracy respectively. From this, we can finally conclude that model with all the features is best fitted the data set and it has given the best possible accuracy of 95.32%. However, the need is that we should train different algorithms and then analyze their prediction for a continuous value. By improvisation of errors, this work could be used for different cities.

7. References

[1] Maharshi Modi, Ayush Sharma, Dr. P. Madhavan, Applied Research on House Price Prediction Using Diverse Machine Learning Techniques, International Journal of Scientific & Technology Research VOLUME 9, ISSUE 04, APRIL 2020.
[2] Chen Jie and Hoogh K D 2019 A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide, *in Environmental International*. vol.130.

[3] Araki S and Shima M 2018 Spatiotemporal land use random forest model for estimating metropolitan no 2 exposure, *in Japan Sci. Total Environ.* pp. 1269-1277.

[4] Nissan Pow and Janulewicz Emil and L. Liu. 2014 Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal.

[5] Byconghwa Park and Jae Kwon Bae 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications* 2928-2934.

[6] Rahadi R A and Wiryono S K and Koesrindartotoor D P and I.B. Syamwil 2015 Factors influencing the price of housing in Indonesia, *Int. J. Hous. Mark. Anal.*, vol. 8, no. 2, pp. 169-188.

[7] Case Study: Malang, East Java, Indonesia Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization.

[8] Ferwerda J and Hainmueller J and Hazlett C J 2017 Kernel-based regularized least squares in r (krls) and stata (krls). J. Stat. Softw.79.