An Empirical Study Of Machine Learning Regression Models to Predict Health Insurance Cost

[1] Y. Angeline Christobel , [2] Suresh Subramanian

[1]Dean, School of Computational Studies, Hindustan College of Arts & Science, Chennai

[2] Chairperson, Department of Multimedia, College of IT, Ahlia University, Kingdom of Bahrain.

Abstract

Medical expenses are a substantial ongoing expense in human existence. When insurance members become ill or have an accident, health insurance firms guarantee that their health costs will be covered. This paper examines personal health data in order to estimate insurance premiums for people. The effectiveness of several machine learning regression algorithms in estimating the cost of medical insurance is reviewed, as well as the factors that influence the cost of health insurance. Linear, Ridge, Lasso, and Polynomial regression techniques were employed in the investigation. The results of the experiments demonstrate that polynomial regression is the best method for forecasting the cost of health insurance.

Keywords: Regression, Linear, Ridge, Lasso, Polynomial

1. INTRODUCTION

The most pressing issue in today's globe is healthcare spending. The World Health Organization (WHO) estimates that global healthcare spending is at 7.5 trillion US dollars[1]. The reason for this exorbitant cost is because health care has a low level of accountability. Patients are overcharged, and needless procedures or medications are administered to them. The challenges of accountability can be solved if healthcare costs can be predicted with great precision. Nowadays, health insurance is required to cover medical expenses, and practically everyone is affiliated with an insurance provider. Many experts and practitioners believe that health insurance is an essential part of the medical field's infrastructure. Medical costs, on the other hand, are impossible to forecast because the large bulk of funds come from patients who suffer from rare diseases. In the prediction phase, many machine learning approaches are used. The elements that influence the cost of health insurance vary from one organization to the next. This research aids in estimating how much a person's health insurance will cost. Early health insurance projection informs consumers about the amount of money they will receive from the insurance provider as well as the benefits they will receive. In the field of information technology, machine learning is one of the fastest-growing fields. It aids data scientists in the discovery of hidden patterns in massive datasets. Regression analysis is a type of supervised machine learning that aids in the prediction of data. It's a crucial tool for data modeling and analysis. It examines the effects of variables evaluated on several scales, allowing researchers to determine the optimal set of variables to utilize when developing prediction models. By comparing regression analysis approaches such as Linear, Ridge, Lasso, and Polynomial, this article focuses on regression analysis to estimate the insurance amount. Polynomial regression provides good accuracy for estimating insurance payments, according to the findings.

Researchers employ machine learning techniques to forecast the cost of health care. B. Nithya et al. used machine learning models to predict various forms of cancer sickness in [2]. For a medical prediction system, Anuja Tike[3] et al. used hierarchical decision trees. Their results reveal that their price prediction algorithm has a high level of accuracy. To forecast Intensive Care Unit (ICU) costs, Moran et al. [4] used linear regression approaches. For the analysis of medical expenses in the healthcare system, Gregori [5] et al. used a variety of regression approaches like“ Multiple Linear Regression”, “Decision Tree Regression, and “Gradient Boosting”. Decision Tree Regression was evaluated to forecast insurance costs by Nidhi Bhardwaj et al. [7], and the Gradient Boosting Decision Tree technique was shown to be superior to the other algorithms. In their study, Decision Tree Regression came in the first place, and they discovered that it is a suitable method for predicting runoff time on sensor nodes. Imran Chowdhury Dipto et al. examined “Logistic Regression”, “Support Vector Machine”, and “Artificial Neural Network” to predict coronary artery disease in [10], and found that among the trained Algorithms, “Artificial Neural Network” has the highest accuracy and “Logistic Regression” has the lowest. SH Kok [11] et al. used a distributed denial-of-service intrusion to test various machine learning algorithms and discovered that Random Forest provides the best accuracy. Machine learning techniques have been widely used to predict healthcare costs, but the data that is used varies. For example, the Japanese Public Health Insurance Database [12] and France's state-wide claims database [13] are used in machine learning applications for predicting individual healthcare expenditures.

The following is a breakdown of the paper's structure: The methods will be explained in section 2. The experimental data and comments are presented in Section 3. Section 4 will contain the conclusion.

2. METHODOLOGY

Figure 1 depicts the study's proposed architecture. The dataset “insurance” is taken from Kaggle [6]. There are 6 attributes and 1338 rows in this dataset. 'Age, gender, bmi, children, smoker, and costs' are the attributes. This dataset contains information about the patient, as well as the total medical expenses charged to the plan for a calendar year.' Expenses' is a dependent variable, whereas the others are independent variables. Duplicate values are deleted during the data pre-processing step. A label is a name given to each value in a category column. To transform the categorical value to numerical value, the label encoding technique is utilized. Training accounts for 80% of the dataset, whereas testing accounts for 20%.

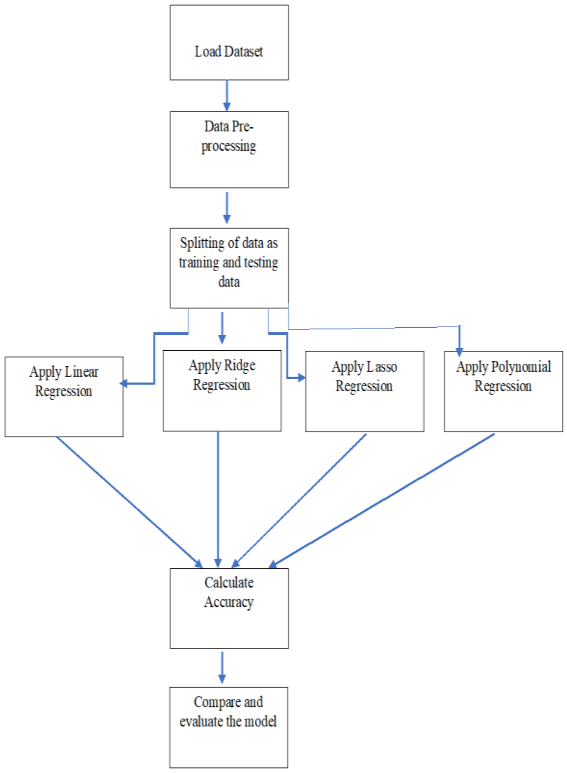


Figure 1: Proposed Architecture

2.1 DATA INTERPRETATION

Figure 2 shows the density distribution of expenses. Skewness is a metric that assesses the degree of symmetry, or lack thereof, in a particular circumstance. If a distribution or data set appears the same on all sides of the graph to the left and right of the center point, it is said to be symmetric. Kurtosis is a measure of how heavy-tailed or light-tailed the data are when compared to the normal distribution, according to the normal distribution. Heavy tails or outliers are more probable in data sets with a high kurtosis than data sets with a low kurtosis. When the kurtosis of a data collection is low, it is more likely that there will be no outliers [14]. The most extreme instance would be if there is a uniform distribution. There may be a few adjustments that are outliers, but we can't proclaim the figure to be an outlier because there may have been instances where the cost of medical treatment was truly relatively cheap.

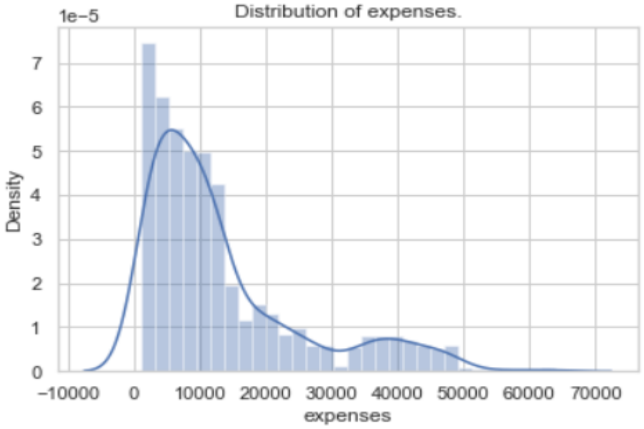


Figure 2: Density distribution of expenses

Figure 3 depicts the cost of living for smokers and non-smokers as a function of age. Smokers are represented by red dots, whereas non-smokers are represented by blue dots.

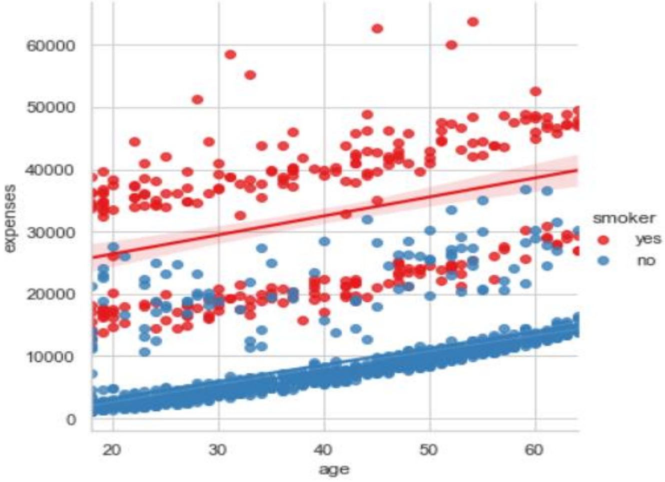


Figure 3: Age Vs Expenses

Expenses vs. BMI are depicted in Figure 4. The x-axis displays the BMI readings, while the y-axis displays the costs. It can be seen that when the BMI values are changed, the insurance premiums change as well.

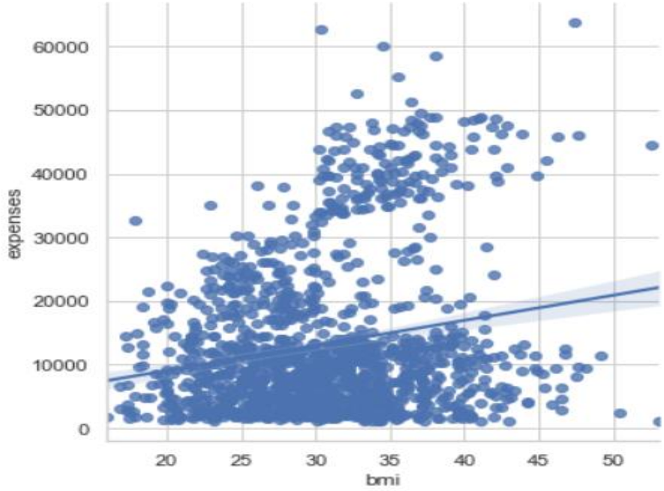


Figure4: BMI vs Expenses

Figure 5 depicts the costs by region. The region is shown on the x-axis, while the expenses are shown on the y-axis. In comparison to other regions, the southwest has higher costs.

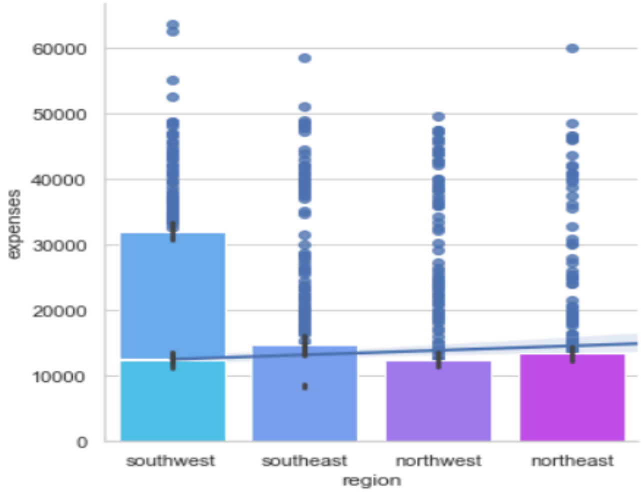


Figure 5: Region vs Expenses

2.2 REGRESSION MODELS

The link between the outcome and the relevant factors is quantified using regression analysis. “Linear Regression”, “Ridge Regression”, “Lasso Regression”, and “Polynomial Regression” are all investigated and compared in this study.

2.2.1 LINEAR REGRESSION

A statistical model is a linear regression. It looks at how a dependent variable interacts with a group of independent variables. When the value of the independent variable changes, the value of the dependent variable changes as well. The following is the relationship's equation:

$$Y = A + B*X$$

Where
X → Input variable (training data)

B → Coefficient of X

A → Intercept (Constant)

Y → Predicted Value (Calculated from A, B and X)

2.2.2 RIDGE REGRESSION

Ridge regression is a linear regression extension. In this model, the loss function is changed to reduce the complexity of Ridge regression. A penalty parameter is introduced that is equal to the square of the coefficients' magnitude. The following is the Ridge Regression equation:

Loss function = OLS + alpha * summation (squared coefficient values)

The parameter to be picked in the above loss formula is alpha. Over-fitting is possible with a low alpha value, whereas under-fitting is possible with a high alpha number.

2.2.3 LASSO REGRESSION

A modification of linear regression is the Least Absolute Shrinkage and Selection Operator (Lasso). The loss function in Lasso is changed to reduce the model's complexity by limiting the sum of the absolute values of the model coefficients. The following is the Lasso Regression loss function:

Loss function = OLS + alpha * summation (absolute values of the magnitude of the coefficients)

The penalty parameter to choose in the above loss function is alpha. When using an L1 norm constraint, some weight values are forced to zero in order for other coefficients to take non-zero values. When a dataset has a lot of dimensionality and correlation, Lasso regression is utilized.

2.2.4 POLYNOMIAL REGRESSION

When the link between independent and dependent variables is modeled as the nth degree of a polynomial, Polynomial Regression is utilized. The non-linear relationship between the independent variable, x, and the dependent variable, y, is fit by polynomial regression. The following is the equation for Polynomial Regression:

$h(x) = Q0 + Q1*x + Q2*x^2 +\ldots Qn*x$^n

When the data points are far away from the line and the error rate is large, simple linear regression fails to generate the best fit line for the data. The line or curve created using polynomial regression is optimal because it fits nearly all of the data points and has a low error rate. When two variables have a non-linear relationship, polynomial regression is utilized.

The analysis of the four regression models and the metrics used is given in the next section

4. EXPERIMENTAL RESULTS AND DISCUSSION

Root Mean Squared Error (RMSE), R-Squared(R^2), and accuracy are the metrics used to analyze regression models. The average of the square of the difference between the anticipated and original values is the Mean Squared Error (MSE). The following is the formula for determining MSE:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t' - y_t)^2$$

Where

y_t' is the predicted value,
y_t is the original value, and
n is the total number of values in test set
.

Root Mean Squared Error(RMSE) is the square root of the MSE. The formula for calculating RMSE is

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t' - y_t)^2}$$

Where

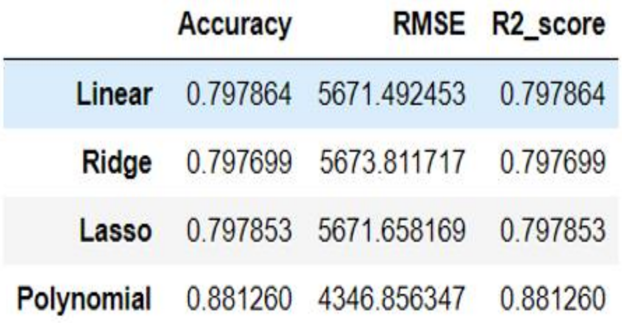
y_t' is the predicted value,
y_t is the original value, and
n is the total number of values in test set

R-squared(R^2) is also known as the coefficient of determination. It is a statistical measure. It finds how close the data are to the fitted regression line. The formula for R^2 is

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

The performance score of RMSE and R^2 value of all regression modelsare summarized in Table 1.

Table 1: Outcome of Regression models



	Accuracy	RMSE	R2_score
Linear	0.797864	5671.492453	0.797864
Ridge	0.797699	5673.811717	0.797699
Lasso	0.797853	5671.658169	0.797853
Polynomial	0.881260	4346.856347	0.881260

Table 1 shows the highest R^2 score and lowest RMSE for Polynomial Regression. The comparison of the accuracy of all models is shown graphically in Figure 6.

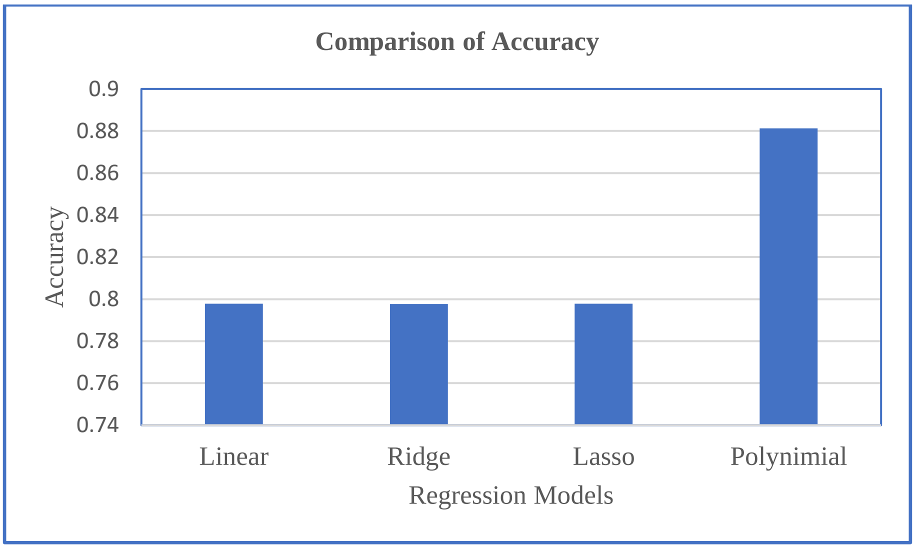


Figure 6: Comparison of Accuracy

Table 1 and Figure 6 show that polynomial regression has higher accuracy, highest R^2 score, and lowest RMSE for forecasting insurance amounts than the other regression techniques. Because the ridge and lasso techniques are extensions of linear regression and work according to the distribution of data in the dataset, they have nearly the same accuracy as linear regression.

4. CONCLUSION

In machine learning, choosing a model is dependent on the dataset, and it's critical to understand how the dataset's properties interact with one another. People are having difficulty paying their medical fees these days due to several types of viruses. So, if the cost of health insurance is known ahead of time, it will make it easier for them to pay the bill. In the field of prediction, regression techniques are quite important. In this paper, four regression analysis algorithms were compared to predict the insurance amount: “Linear Regression”, “Ridge

Regression", "Lasso Regression", and "Polynomial Regression". It was discovered that Polynomial Regression is the best, with an accuracy of 88 percent, the highest R^2 score, and the lowest RMSE.

5. REFERENCES

1. Organization, W.H. Public Spending on Health: A Closer Look at Global Trends; Technical report; World Health Organization: Geneva, Switzerland, 2018.
2. B. Nithya, Dr. V. Ilango, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems ICICCS 2017, 978-1-5386-2745-7/17/$31.00 ©2017 IEEE.
3. A. Tike and S. Tavarageri. (2017). A Medical Price Prediction System using Hierarchical Decision Trees. In: IEEE Big Data Conference 2017. IEEE, 978-1-5386- 2715-0/17/$31.00 ©2017 IEEE.
4. Lahiri and N. Agarwal, "Predicting healthcare expenditure increase for an individual frommedicare data," in Proceedings of the ACM SIGKDD Workshop on Health Informatics, 2014.
5. Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti, and E. Pagano, "Regression modelsforanalyzing costs and their determinants in health care: an introductory review," International Journal for Quality in Health Care, vol. 23, no. 3, pp. 331–341, 2011.
6. https://www.kaggle.com/mirichoi0218/insurance
7. Nidhi Bhardwaj , Rishabh Anand, "Health Insurance Amount Prediction", International Journal of Engineering Research & Technology, Vol. 9 Issue 05,,pp.1008-1011,2020.
8. A. Lakshmanarao, Chandra Sekhar Koppireddy, G.Vijay Kumar,"Prediction of medical costs using regression algorithms" , Journal of Information and Computational Science, Volume 10 Issue 5 – 2020
9. Marwan Khan, Sanam Noor, "Performance Analysis of Regression-Machine Learning Algorithms for Predication of Runoff Time", Agrotechnology, Vol. 8 Iss. 1 No: 187,2019
10. Imran Chowdhury Dipto, TanzilaIslam,H M Mostafizur Rahman, Md Ashiqur Rahman," Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease" Journal of Data Analysis and Information Processing, ,8, 41-68, 2020
11. SH Kok, Azween Abdullah, Mahadevan Supramaniam, Thulasyammal Ramiah Pillai, Ibrahim Abaker Targio Hashem, " A Comparison of Various Machine Learning Algorithms in a Distributed Denial of Service Intrusion", International Journal of Engineering Research and Technology, Volume 12, Number 1, pp. 1-7,2019
12. Y. Nomura, Y. Ishii, Y. Chiba et al., "Does last year's cost predict the present cost? An application of machine leaning for the Japanese area-basis public health insurance database," International Journal of Environmental Research and Public Health, vol. 18, no. 2, 2021.
13. A. Vimont, H. Leleu, and I. Durand-Zaleski, "Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France," The European Journal of Health Economics, pp. 1–13, 2021.
14. D. B. Madan and K. Wang, "Option implied VIX, skew and kurtosis term structures," International Journal of Theoretical and Applied Finance, vol. 24, no. 5, Article ID 2150030, 2021.