IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

REGRESSION ANALYSIS AND PREDICTION OF MEDICAL INSURANCE COST

Ayushi Bharti*[1], Lokesh Malik[2]

[1,]U.G.Scholar, [2]P.G.Scholar,[1,] Electronics and communication engineering,[2]Computer Science and Engineering, [1]MSIT,[2] DU, New Delhi, India

ABSTRACT

Many elements affect the expenses of health insurance and it's far quite a tough project to analyze the sample from those capabilities. We use a regression version to recognize and study a complex sample that enables us to predict the price of medical insurance.

In this paper, we used a dataset from Kaggle that consists of 6 capabilities and 1338 instances.

We will be taking the assistance of diverse regression models and reading which model's overall performance is excessive in predicting medical insurance.

Regression is a statistical procedure for calculating the cost of a primarily based variable from an impartial variable. Regression measures the affiliation among variables. it's far a modeling method wherein a based variable is predicted primarily based on one or more unbiased variables. Regression evaluation is the maximum widely used of all statistical techniques. This text explains the primary concepts and explains how we can do regression calculations.

INTRODUCTION

Everyone's existence revolves around their health. Good health refers to a person's capability to deal with the surroundings on a physical, emotional, mental, and social level. Everyone's existence is generally best unless some sort of health hassle arises that is uncertain and can not be expected before it happens. needs which include the desire of owning a house or a car or some other device of social popularity or different consumer durables of comfort can be postponed if the circle of relatives has inadequate savings and poor assets of income. However, this is not the case with the unexpected medical feature which wishes immediate money, useful resources and impacts the financial savings of the own family. economic stress on medical grounds can certainly smash the long-term monetary dreams of a family which may consist of education or marriage of children and retirement plans besides goals stated supra. One can also wonder about a solution to conquer this sort of crisis and the solution to that is none apart from health insurance which will assist in the protection of the good health of an individual and their own family without growing any possibility of a monetary crisis and disturbing monetary stability. Health insurance is a product of preferred coverage that covers fees associated with the medicine and surgical procedure of an insured which could be an individual, family, or a collection of people. It's an association in which a person, circle of relatives, or a group purchases medical health care coverage in advance with the aid of payment of a fee known as a top class. In other phrases, health insurance is an arrangement that enables to delay, defer, lessen or avoid charges related to the medical expenses of an insured. The insurer will either make certain cashless remedies of medical ailments or offer repayment of medical costs incurred under the policy in any of the network hospitals throughout the country.

APPROACH

The goal of this paper is to tell in advance about the health insurance expenses, which could prove to be very beneficial for insurers and patients and this can be achieved by managing assets and selecting appropriate plans. To fulfill this goal, we have already discussed above various techniques of Machine Learning which can be implemented to the dataset, and the analysis of the dataset is also shown in this paper. The attributes which play an important part in getting higher precision, are also told in this paper. And because of this, a patient need not be tested on all attributes, but only on the ones which play a major role, and thus the expense of the patient can be minimized.

Data Source

In this study we have used the Kaggle dataset of Medical Insurance cost and trained our model on attributes like age, sex, BMI, children. smoker, region which contains 1070 training instances and 267 testing instances.

You can see all the attributes in the following table:

Table 1: Attribute table

S.NO.	**Attribute**	**Remarks**
1	Age	Age of primary beneficiary
2	Sex	Gender of patients; Female(0) and Male(1)
3	BMI	Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
4	Children	Number of children covered by medical insurance.
5	Smoker	Smoking
6	Region	The beneficiary, residential area in the US, North-East, South-East, South-West, and North-West.
7.	Charges	Medical Insurance price

Total of 5 algorithms are used in this research to predict medical insurance cost and then a model is created with maximum possible accuracy.

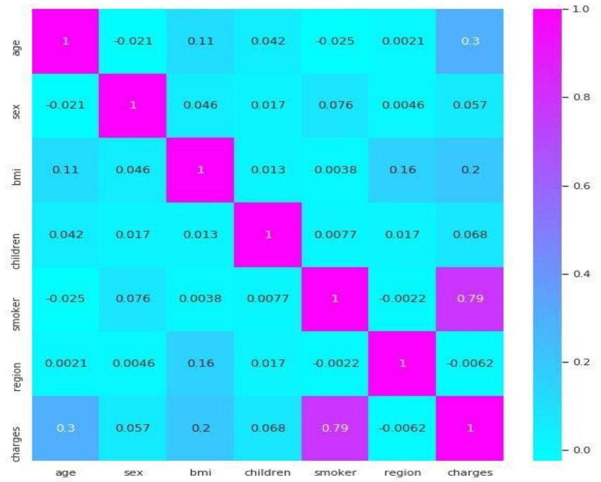

Figure no.1: Correlation plot of features

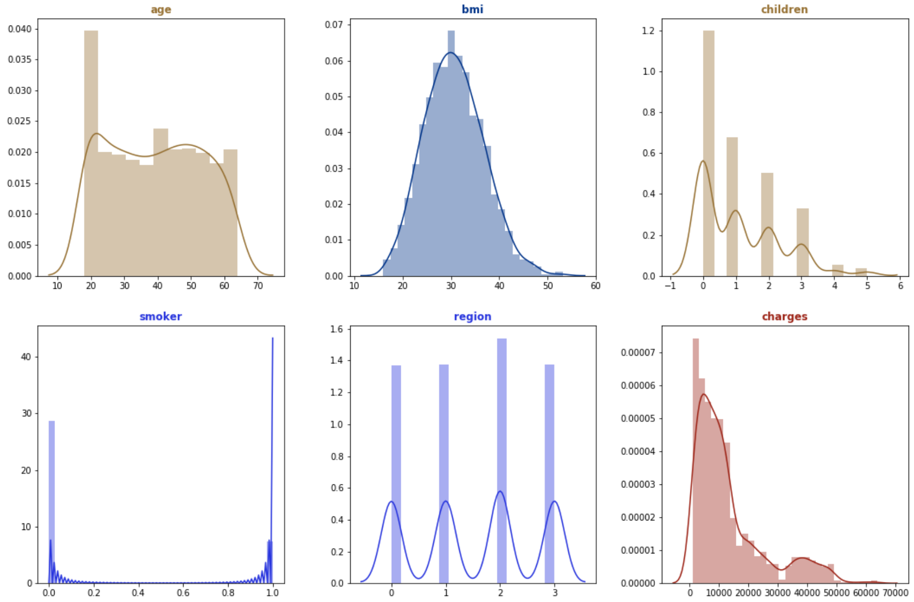

Figure no.2: Distribution of features

Data Pre-Processing

The real-life records include quite a few noisy data and missing values data. So, to make accurate predictions, we first pre-process the data to overcome the problems which might come from missing values and noisy data. We also have made a flow graph of our proposed model, which can be seen below:

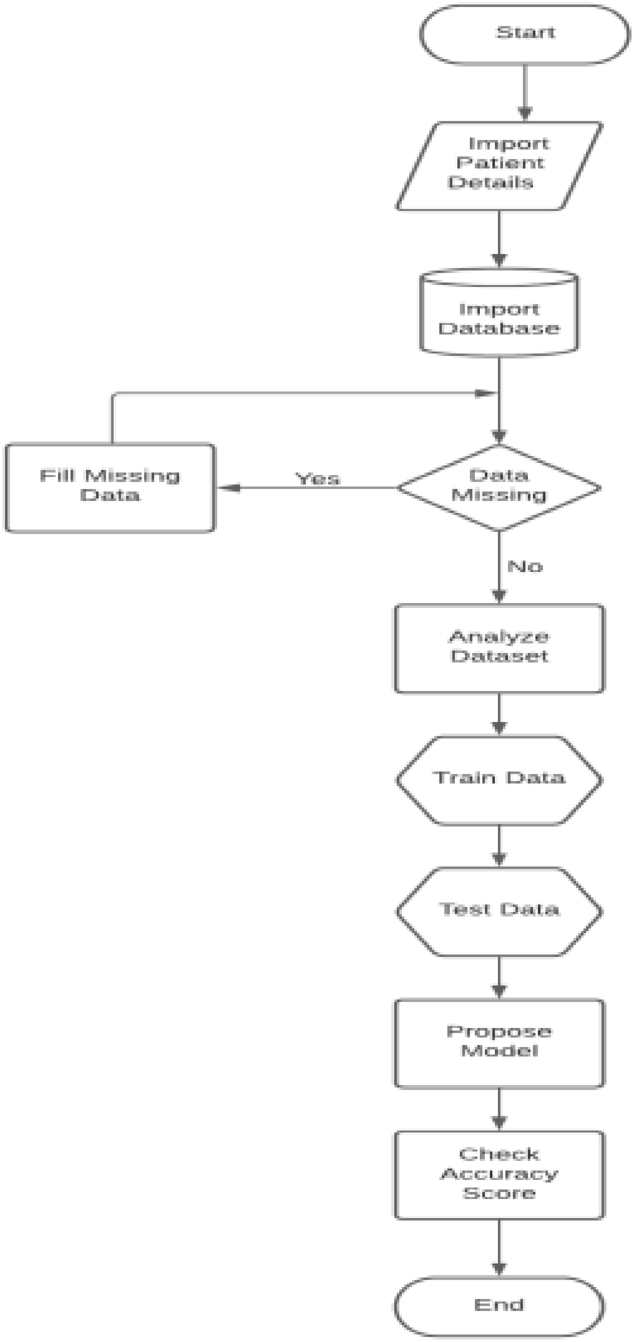

Figure 3: Model Flowchart

Step 1 clean: To get correct results, the information must be cleaned and missing values need to be filled.
Step 2 remodel: on this, we use smoothing, aggregation, and normalization tasks to make information more comprehensible through converting the layout of the data.
Step 3 Integration: before processing, we need to integrate the data because it is probably obtained from diverse sources, not always from a single one.
Step 4 reduction: The acquired data is very complicated and needs to be formatted for accomplishing preferred outcomes.
The data is then classified and divided into test and training data which can be then attempted on different algorithms to get accuracy score results.

Algorithms Used

Linear Regression

Linear Regression is the simplest supervised machine learning algorithm where the target output is continuous. Linear Regression is created to predict values within a continuous range, rather than trying to classify different classes.
Behind the algorithm, it uses the traditional slope-intercept form, where m and b are the variables of our algorithm. It tries to predict the best fit line that describes the relationship between the input feature and the target variable.

Ridge Regression

Ridge regressor is an updated version of Linear regression that adds a regularized term in the equation of Linear regression. It helps the cost function of the linear regressor that forces the learning algorithm to fit the data accurately. We can also control the regularized term by constant named alpha, it helps cost function in reducing the variance of the estimates.
We can also control the bias and variance by tuning the alpha parameter. If we increase alpha the biases in the model will increase and variance will decrease.
It shrinks the input features and can be used to prevent multicollinearity and reduces the model complexity by coefficient shrinkage.

Lasso Regression

Lasso regression is another updated version of linear regressor and a famous L1 regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. It can be used as an automatic feature selection model to select important features in a model.

Elastic Net Regression

Elastic regression is a combination of L1 regularization and L2 regularization strategies. It provides regularization penalties to the loss function during training. It penalized linear regression on the sum of squared coefficients as well as on the sum of the absolute coefficients to control multicollinearity and feature selection.

Random Forest Regressor

Random forest regressor creates a forest of decision trees. In a random forest, regressor output no longer relies upon one individual tree. every tree in the forest will give its output class and the class with the highest votes will be the final output.

It is a bagging model which uses bootstrap sampling to train the model and is typically used to reduce the variance of models.

Random forest algorithms can be used for both regression and classification and we also get the satisfactory result when we use a random forest regression model.

RESULTS & ANALYSIS

The objective of this study was to predict the price of medical insurance. We used various regression techniques like Linear Regression, Ridge Regression, Lasso Regression, Random forest, and Elastic Net. For this experiment, we recommend that one should use a machine with at least 16GB RAM and generation of Intel Processor shall be at-least 9th technology or above. Dataset was split into 2 sets, one set for training and one for testing purposes. We used python programming for noting the results of carried-out techniques on test and training datasets.

Table no.2: r2-score

S.no.	**Algorithm**	**r2-score**
1.	**Linear Regression**	**0.7699**
2.	**Ridge Regression**	**0.7996**
3.	**Lasso Regression**	**0.7999**
4.	**Random Forest Regression**	**0.85388**
6.	**Elastic Net**	**0.5519**

CONCLUSION

The main objective was to explain distinct regression strategies of data mining that may prove to be useful in predicting effective medical health insurance price and facilitates insurers and patients to obtain their goals. Our aim turned into getting as high accuracy with a decreased range of attributes as possible.

We used 7 essential attributes and used five regression techniques particularly Linear Regression, Ridge Regression, Lasso Regression, Random forest, and Elastic Net. The data was used in the model after pre-processing it. The best rating was received by using Random forest. In the future, we can practice more strategies to get even more accurate scores for medical insurance costs.

REFERENCES

[1]. Medical cost personal datasets insurance forecast by using Linear Regression https://www.kaggle.com/mirichoi0218/insurance

[2]. Linear regression analysis study from researchgate https://www.researchgate.net/publication/324944461_Linear_regression_analysis_study

[3]. Understanding of lasso regression from My great learning https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#11

[4]. What is ridge regression from My great learning https://www.mygreatlearning.com/blog/what-is-ridge-regression/

[5].Linear regression in machine learning from great learning https://www.mygreatlearning.com/blog/linear-regression-in-machine-learning/

[6].Machine learning ridge regression using sklearn from geeksforgeeks https://www.geeksforgeeks.org/ml-ridge-regressor-using-sklearn/

[7]. Five types of health insurance plan from acko articles https://www.acko.com/articles/health-insurance/5-types-of-health-insurance-plan-in-india/

[8]. K Swathi and R Anuradha (2017), Health insurance in India- An overview.

[9]. Suman Devi and Dr. Vazir Singh Nehra (2015), The problems with health insurance sector in India.12. Shatakshi Chatterjee, Dr. ArunangshuGiri, Dr. S.N. Bandyopadhyay (2018), Health insurance sector in India: A study.

[10]. Types of health insurance from reliance general https://www.reliancegeneral.co.in/Insurance/Knowledge-Center/Insurance-Reads/Types-Of-Health-Insurance-Covers.Aspx