Health Insurance Cost Prediction Using Machine Learning Technique

Rohit Amrutkar[1], Abhishek Chikurdekar[2], Atharv Salunke[3], AbhijeetWaghmare[4],Minal Bodke[5]
[1,2,3,4]*Department of Computer Engineering, PCCOER, Ravet, Pune, India*
[5]*Department of COMPUTER, PCCOER, Ravet, Pune, India*
E-mail: abhishekchikurdekar802@gmail.com

Abstract-Insurance is a policy that diminish or eradicate the expenditure loss appear due to various risk. Different factors may have an impact on the Insurance cost. Machine learning is a field of study that gives computers potential to adapt without human interference. Machine learning is widely used in insurance industry. Machine learning allow insurance companies to acknowledge the customers in a better way and create an insurance cover that victual to their needs and profile. In this paper, the propose system will examine the individuals health details to forecast the health insurance amount prediction and exhibit the insurance plans. Four regression models such as Linear Regression, Support Vector Regressor, Random Forest Regressor and Gradient Boosting have been implemented and their performance is measured using Mean Absolute Error, Root Mean Absolute Error and Coefficient of determination. Based on the analysis, it is found that the Gradient Boosting performs better than the other regression algorithms.

***Keywords*: Health Care, Regression, Cost prediction, Health Insurance**

I. INTRODUCTION

In recent years, staying healthy is extremely important but several factors such as extreme workload, lack of exercise, unhealthy personal habits, and unhealthy food can have an undesirable effect on the health of most people. The consequence could be multiple health problems and immoderate cost of medical treatment. But, health problems cannot usually be avoided, so the financial activities supersector has developed numerous products to help people during medical emergency [2].

Therefore, A health insurance plan can be a solution to cover the rising medical costs. It provides financial protection by covering the costs related to treatment, hospitalization, free health check-up, and pre and post hospitalization expenses.

Nowadays, a health insurance policy is a essential due to the extensive benefits it offers. But, most of the people are unaware of it and they do not invest money in health insurance at all. In some cases, people are fooled by insurance agent and they may unnecessarily buy some expensive health insurance[3].

Cost of insurance varies according to different attributes. For a variety of people accurately predicting individual medical treatment costs is critical. Accurate cost evaluation can help person to plan for the future. Furthermore, knowing ahead of time what their probable expenses for the future can assist patients to choose insurance plans with appropriate premiums according to their financial condition.

Conventional programming goes to manually written program that uses input data and runs on a computer to generate output. In Machine Learning the input data and output are add to an algorithm to create a program. Machine Learning is an application of Artificial Intelligence that gives systems the ability to automatically learn and improve from experience without being externally programmed [8]. It is one of the Artificial Intelligence algorithms which is developed to copy human intelligence.

Machine Learning can be categorized into three different types. These types are supervised machine learning (a task-driven approach) used for classification/regression and all data labeled; unsupervised machine learning (a data-driven approach) used for clustering and all data unlabeled; and reinforcement learning (learning from mistakes) used for decision making.

Nowadays, ML has become well liked in research and is used in large number of applications such as Image processing, Data mining, Speech Recognition, text mining, social network analysis, health sector and so on [10] [11] [12].

In personalized digital marketing machine learning plays an significant role like Showing relevant Ads to customers, Ads click prediction, identifying target customers, etc. are some of the significant applications of ML in marketing sector.

In Banking and Finance, machine learning helps financial sector to protect from money laundering, fraud, identifying valuable customers, illegal

financial detection, etc. It also helps financial companies with stock market prediction and demand forecasting.

In automobile sector almost every automobile industry is using Artificial Intelligence for breakdown prediction, fuel consumption and self-driving.In the insurance sector, Machine Learning can help claim processing automation, more extensive product choice. In healthcare, Machine Learning algorithms are particularly good at accurate insurance case probability calculation and cost setting.

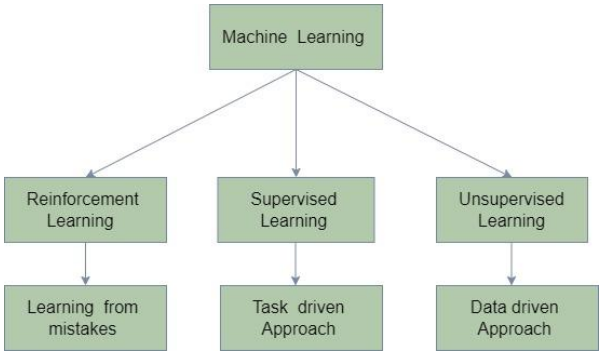

Fig.1: Types of Machine Learning

II. LITERATURE REVIEW

The literature survey shows the comparative study of various machine learning algorithms to predict the health insurance cost.

In 2020, A. Lakshmanarao et al, proposed machine learning model for predicting medical costs. They applied support vector regression, multiple linear regression, random forest regression and decision tree regression. They also applied MLR with backward elimination technique and observed that bmi, age are features which decides the dependent variable. Out of all experiments random forest given better results [1].In 2021, Mohamed Hanafy et al, proposed various machine learning regression models and deep neural networks to predict charges of health insurance based on specific attributes. Prediction of insurance cost based on certain factorswill help the insurance policy provider's to attract consumers and save time in formulating plans for every individual [2]. In 2020, Nidhi Bhardwaj et al, proposed three regression models evaluated for individual health insurance data. The health insurance data was used to develop the three regression models, and the predicted premiums from these models were compared with the actual premiums to compare the accuracy of these models. Various factors were used andthere effect on predicted amount was examined. It was observed that person's age and smoking status affects the prediction most in every algorithm[3]. In 2020, Shinde et al, applied three regressions techniques. They made combinations of attributes along with ML algorithms to achieve better accuracy to predict medical cost. The proposed system examined many regression models and neural network models like support vector machine,multiple linear Regression ,XG boost, random forest regressor and Deep neural network.The Deep neural network was choosen as best technique with RMSE value of 0.0695 and accuracy of 87.95. [4] Mladenovic et al (2020) used Artificial neural Network (ANFIS) for evaluation of medical insurance cost with RMSE score of 7464.631.[5] Tkachenko et al ., (2018) suggested a methodology for health insurance cost prediction which was costwise Linear technique using SGTM neural-like structure. According to their observations the proposed method achieved good performance with MAPE percentage of 30.60 and MAE value of 3453.29. [6] In 2021, Krittika Dutta et al, the predictive models that are described have been used for calculating the cost of the health policy as accurately as possible, which will be supposedly very much beneficial for the health care organization in establishing better medical facilities. They utilised different regression algorithms like random forest, decision tree regression, linear regression and polynomial regression. By comparing all the results, best model was found to be the random forest [7]. In 2018, BattaMaheshet al, proposed various machine learning algorithms. The result shows that unsupervised learning gives optimised results for larger dataset [8].

Table1: Attributes of the Dataset

Attributes	Description
Age	Users age
Sex	Gender
No of children	Number of children
B. M. I	Person's body mass index
Region	Residential area of the person
Smoker	To know if person smokes or not
Annual Income	Salary of person
Charges	Individual medication cost

III. PROPOSED METHODOLOGY

3.1 Dataset

Data set is taken from Kaggle [9] in csv format. The Dataset contains continuous as well as categorical data. The dataset has 1338 rows and 8 premium forecast attributes such as age, sex, no of children, body mass index, region, smoker, annual income and charges. The age attribute defines the user's age, sex defines male or female, no of children to get the count of children, body mass index describes person's body mass index such as height and weight. The region attributes helps to decide a particular person is from which regional area, smoker attributes to know if a person smoke or not, annual income to ask for annual salary of a person and charges defines the individual medication amount. The following Table1 will give the glimpse of the attributes.

3.2 Pre-processing

Once, we get the raw data it need to be pre-processed. The raw data could be in any of the structured, semi-structured and non-structured format. If the raw data contains some corrupted or incomplete values it might lead to incorrect result. So, pre-processing is considered as one of the important process in resolving machine learning problem. The dataset is pre-processed using data cleaning techniques such as removing missing values using mean, median and mode.Pre-processing on the dataset file is done to get the correct accuracy on the model.

3.3 Feature Extraction

If the raw data contains categorical values the dataset will not be trained. Feature Extraction is the method to transform the categorical values into integer values for training the model. The dataset file need to haveinteger values to train the model. In dataset file thecolumns such as sex, region and smoker contains categorical values , this columns are transformed and integrated using feature extraction method. The sex column is transformed such as male will be considered as 1 and female will be considered as 0. For smoker it will be considered as if smoker says yes it will be 1 and if no it will be 0 and same goes for the region column also. For this there is method called Label Encoder is used.Fig 2 represents the planfor the health amount prediction

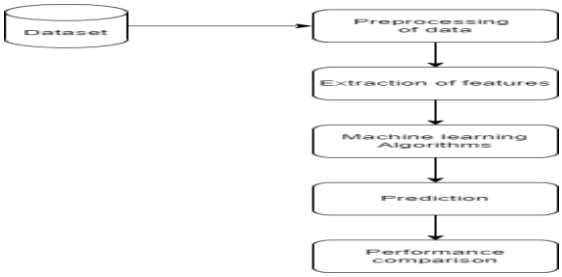

Fig.2: Execution Plan

3.4. Machine Learning Algoritham Used

3.4.1 Multiple Linear Regression

Multiple linear regression is an statistical approach.It is used for predictive analysis.Multiplelinear regression is called an enlarged version of simple linear regression. It comes in picture when there is a needto forecast a single outputbased on more than 2 inputs. The forecasted variable is known astarget variable and the variables used for forecasting target variable are known as regressor variables[3].In multiple linear regression the target variable should be continuous, but the regressor variable could be in continuous or categorical form. The attributes in the dataset should model the linear equation with the target variable. The main aim of multiple linear regression is to create a line through n – dimensional space data points. For example, thedataset file contains 8 attributes out of which 7 attributes are independent attributes (age, sex, bmi, no of children, region, smoker and annual income) and target attribute is the charges attribute. The multiple linear regression aim to create a linear relation with the independent attributes and to find the health insurance forecast.

Following is the equation of the multiple linear regression algorithms which is used for forecasting the regression problems.

$$y = b_0x_1 + b_1x_2 + b_2x_3 + \dots\dots\dots\dots + b_nx_n$$

where y is the targeted value and x1, x2,…..,xn are the independent values with b0,b1,……,bn as a intercept.

3.4.2 Support Vector Machine

Support vector Machine is one of the widely used algorithm. It comes under Supervised learning approach. It can be used for both to analyse the regression problems as well as classification problems. Support vector machine contains two types such as linear support vector machine and

non-linear support vector machine. In linear if the dataset is linear then it will classify into two classes such that when a new user will be added in the correct class. It can be used to examine non-linear classification using something known as kernel trick which completely plot inputs in the multi-scale feature space[2].It basically draw a line to distinguish between the groups. It is draw in such a way that the gap between the line and the group is more and it will reduce the classification error[8]. The attributes in the dataset will partition the into groups and draw a hyperplane to divide the attributes and reducing the error.

3.4.3 Random Forest

Random Forest algorithm is used for solving both type of problems regression as well as the classification problems. It is build on ensemble learning task comprising of gathering various classifiers to resolve a complicated issue and increase the execution parameter of the model. Random forest requires least training period with respect to the other machine learning algorithms. Random forest forecast the output value with greater accuracy for big and complicated datasets. Another additional use of random forest is that it can give better accuracy even if large amount of data is missing. The random forest algorithm consists of more than one decision tree on several subgroups of the data file and extract the mean to increase the forecasting accuracy of the data file. The more number of trees in forest lead the way to better accuracy on most of the votes and forecast the final output.

3.4.4 Gradient Boosting

Gradient Boosting is based on the boosting methods. The gradient boosting models works for both regression as well as classification problems. This algorithm is the best technique for reducing the loss function of the model for the better execution of the problem. In the gradient boosting algorithm each forecaster rectifies it's forerunner flaw.Gradient Boosting is extemporize on the characteristics of Adaboost to make a powerful and more effective algorithm. Boosting operates on the concept of learning from preceding learner.Boosting operates on back to back connecting the weak learner and clarify the examination which a learner gainright at eachpath.In gradient boosting various decision trees are gathered many weak learner to create a powerful learner. The weak learner are nothing but the individual decision trees. The gradient boosting algorithms operates on back to back connecting due to which it operates slowly , but the accuracy of this algorithm is better than the other algorithms. The loss function is used to control the error between the outcome of the algorithm and the targeted value. The formula which is used for calculating the loss function in gradient boosting is as follows:

$$L = \frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2$$

The formula contains$\hat{y}_i$which describes the forecasted value and y_i describes the noticed value and n shows the number of samples.

Table2:Result Analysis of all the Models

Algorithms	MAE (Mean Absolute Error)	R2_SCORE (Coefficient of Determination)	RMAE (Root Mean Absolute Error)
Linear Regression	41.68442	0.767751	64.563475
Support VectorMachine	82.69873	0.830755	90.938844
Random Forest Regression	26.12709	0.856166	51.114667
Gradient Boosting	24.70355	0.868031	49.702675

3.4.5 Prediction

The proposed system will find the best health insurance forecast for a patient using machine learning models such as Linear Regression, Support Vector Machine, Random Forest and Gradient Boosting. Once, the training and testing of these models is done then it will forecast the medical price for a patient depending on the annual income attribute from the dataset with a great accuracy and this will help the patient to raise a fund for her medical emergencies using insurance policy plans.

3.4.6 Performance

The machine learning models such as Linear regression, Support Vector Machine, Random Forest and Gradient Boosting are implemented and analysed on the different parameters. The parameters are mean absolute error, root mean absolute error and coefficient of determination. The best model will be the one which is having the highest R2 score and minimum mean absolute error.

IV. EXPERIMENTAL RESULT

The implementation of several regression algorithms has been tested for the preference of the premier algorithm , to notice the accuracy and implementation of the model. The implementation of regression algorithms is expressed by analysing on the metrics such as Mean Absolute Error, Coefficient of determination and Root Mean Absolute Error. The implementations of varied metrics shown:

By analysing the varied metrics or parameters of these regression algorithms from the table2 , it was found that the more the coefficient of determination of the model , lesser the mean absolute error of that algorithm. Any model is considered to be best if the coefficient of determination value is higher than the mean absolute error value. By testing the models on this metrics will help us to find the finest algorithm for insurance cost forecasting. Below figures will give the better understanding of the considered metrics.

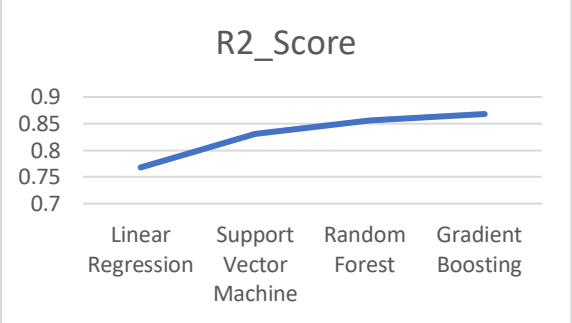

Fig3:Result Analysis based on R2 Score

In the fig3, graph is implemented on the coefficient of determination parameter with respect to the machine learning algorithms. The gradient boosting algorithm is showing higher R2 score as compared to other algorithms with a value 0.868031.

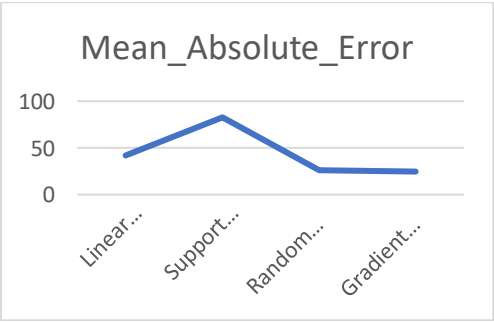

Fig 4: Result Analysis based on Mean Absolute Error

In fig 4 graph is analysed on the mean absolute error parameter and it was observed that support vector machine is with the higher absolute mean error with a value 82.69873 and gradient boosting is with the least one with value 24.70355.

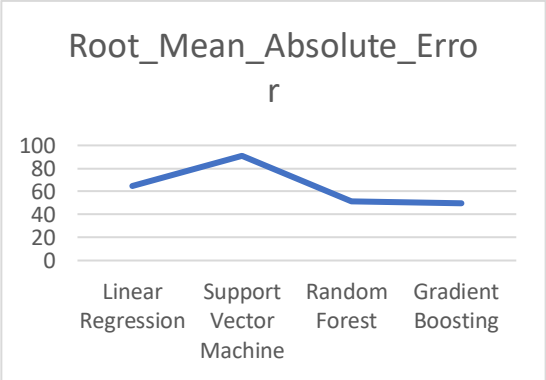

Fig 5: Result Analysis based on Root Mean Absolute Error

The proposed problem is analysed on the root mean absolute error as shown in the fig 6 linear regression is with the value 64.563475, random forest with 51.114667 and gradient boosting is with the least error with a value 49.702675.

By analysing the execution of the metrics or parameters , it was found that the Gradient Boosting Regressor is the finest algorithm among the others with highest r-squared value and lowest mean absolute error value. Gradient Boosting Regressor has 0.868031 r2 score , 24.70355 mean absolute error and 49.702675 root mean absolute error. The Random Forest Regressor attain 0.856166 r2 score , 26.12709 mean absolute error and 51.114667 root mean absolute error. Linear Regressor was the minimal performer with r2 score 0.767751 , 41.68442 mean absolute error and 64.563475 root mean absolute error.

V. CONCLUSION

The regression models which are narrated here have been used for forecasting the amount of insurance policy which will be very helpful for the medical organization for providing better health facilities. The proposed paper have implemented and analysed different machine learning algorithms such as Linear Regression, Support Vector machine, Random Forest and Gradient Boosting. After evaluating all the results it was found that the Gradient Boosting Regressor model is the best model with higher R2 score. Also the annual income attribute will

help to minimize the search for the better policies. The insurance policy forecasting is based on person's own health than the insurance company's conditions. This will help not only people but the insurance company's also to improvise their insurance policies. Future studies can be getting using other deep learning techniques. Improvised methods and techniques of deep learning will be helpful.

6. REFERENCES

[1] A.Lakshmanarao, Chandra Sekhar Koppireddy, G.Vijay Kumar, " Prediction of medical costs using regression algorithms " in Journal of Information and Computational Science, 2020, pp.751-757.

[2] MohemedHanafy, "Predict Health Insurance Cost by using MachineLearning and DNN Regression Models" in InternationalJournal of Innovative Technology and Exploring Engineering (IJITEE) , January 2021, pp.137- 143.

[3] Nidhi Bhardwaj, RishabAnand, "Health Insurance Amount Prediction" in International Journal of Engineering Research & Technology (IJERT), May 2020, pp.1008-1011.

[4] S. C. Autonomous, Misbehaviour detection in C-ITS, vol. 1, no. March.Springer International Publishing, 2019.

[5] S. S. Mladenovic et al., "Identification of the important variables for prediction of individual medical costs billed by health insurance," Technol.Soc., vol. 62, no. August 2019, 2020,doi:10.1016/j.techsoc.2020.101307.

[6]R. Tkachenko, I. Izonin, N. Kryvinska, V.Chopyak, N.Lotoshynska, and D. Danylyuk,"Piecewise-linearapproach for medical insurance costs prediction usingSGTM neural-like structure," CEUR Workshop Proc.,vol. 2255, pp. 170– 179, 2018.

[7] Krittika Dutta, Satish Chandra, Mahendra Kumar Gourisaria, Harshvardhan GM, "A Data Mining basedTarget Regression-Oriented Approach to Modelling of Health Insurance Claims," in Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)IEEE Xplore, 2021,pp.1168-1175.

[8] Batta Mahesh, "Machine Learning Algorithms - A Review," in International Journal of Science and Research (IJSR), 2018,pp.381- 386.

[9] M. Choi, "KaggleDataset:InsuranceCost prediction."https://www.kaggle.com/mirichoi0218/insurance(accessed Jan. 10,2021).

[10] Bhangale, KishorBarasu, and K. Mohanaprasad. "A review on speech processing using machine learning paradigm." International Journal of Speech Technology 24, no. 2 (2021): 367-388.

[11] Bhangale, KishorBarasu, and MohanaprasadKothandaraman. "Survey of Deep Learning Paradigms for Speech Processing." Wireless Personal Communications (2022): 1-37.

[12] Bhangale, Kishor, and K. Mohanaprasad. "Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network." In Futuristic Communication and Network Technologies, pp. 241-250. Springer, Singapore, 2022.