Volume: 10 Issue: XII Month of publication: December 2022

DOI: https://doi.org/10.22214/ijraset.2022.47923

Medical Insurance Cost Prediction using Machine Learning

Mukund Kulkarni[1], Dhammadeep D. Meshram[2], Bhagyesh Patil[3], Rahul More[4], Mridul Sharma[5], Pravin Patange[6]

***Abstract:** Insurance is a policy that helps to cover up all loss or decrease loss in terms of expenses incurred by various risks. A number of variables affect how much insurance costs. These considerations of different factors contribute to the insurance policy cost expression. Machine Learning(ML) in the insurance sector can make insurance more effective. In the domains of computational and applied mathematics the machine learning (ML) is a well-known research area. ML is one of the computational intelligence aspects when it comes to exploitation of historical data that may be addressed in a wide range of applications and systems. There are some limitations in ML so; Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry and thus it requires few more investigation and improvement. Using the machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Decision Tree Regression and Gradient Boosting Regression and also streamlit as a framework. We had used a medical insurance cost dataset that was acquired from the KAGGLE repository for the cost prediction purpose, and machine learning methods are used to show the forecasting of insurance costs by regression model comparing their accuracies.*

***Keywords:** Regression, Neural Networks, Machine Learning, Data Processing*

I. INTRODUCTION

We live on a planet full of threats and uncertainty. Including People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets[1]. But, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to shield them. Therefore Insurance is one of the policies that either decreases or removes loss costs incurred by various risks. The value of insurance in the lives of individuals. That's why it becomes important for insurance companies to be sufficiently precise to measure the amount covered by this specific policy and the insurance charges which must be paid for it. Various parameters or factors play an important role in estimating the insurance charges and Each of these is important. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore very critical to carry out these tasks with high accuracy. So, the possibility of human mistakes are high so insurance agents also use different tools to calculate the insurance premium. And thus ML is beneficial here. ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on insurance data from the past. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This decreases human effort and resources and improves the company's profitability. Thus the accuracy can be improved with ML. Our goal is to predict insurance costs. The value of insurance fees is based on different variables. As a result, insurance fees are continuous. Regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used .[2] Preprocessing of the dataset done first. Then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, Decision Tree Regression and Gradient Boosting Regression. It is found that the gradient boosting provides the highest accuracy with an r-squared value of 86.7853. The inclusion of a novel method of insurance cost estimation is the main goal of this work.

II. DATASET

We had used a dataset from Kaggle Site for creating our prediction model. This data set includes nine attributes and the data set has splitted into two-parts : training data and testing data.For training the model, 80% of total data is used and the rest for testing.To build a predictor model of medical insurance cost the training dataset is applied and to evaluate the regression model, test set is used. The following table shows the Description of the Dataset.

Table I. Dataset overview

Name	Description
Age	Age of client
BMI	Body mass index
No. of kids	Number of children the client have
gender	Male / Female
Smoker	Whether a client is a smoker or not
Alcoholic	Whether a client is drinks alcohol or not
Diabetic	Whether a client is having diabetes or not
region	Whether the client lives in southwest, northwest, southeast or northeast
Charges(Target Variable)	Medical Cost the client pay

III. DATA PREPROCESSING

The dataset includes nine variables, as shown in table 1.[3] From these variables each one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms.

Now the categorical variables are converted into numeric or binary values to represent either 0 or 1. For example, instead of "SEX" with males or females, the "Male" variable would be considered as false (0) if the person is male. And "female" would be (1) see table II; following this phase now, we can apply this data to all regression models used in this study.

Table II: categorical variables after translated into numeric or binary values

Name	Description
Age	Age of client
BMI	Body mass index
No. of kids	Number of children the client have
gender	Male / Female 0=Male 1=Female
Smoker	Whether or not a client smokest 0=yes 1=no
Alcoholic	Whether a client drinks alcohol or not 0=yes 1=no
Diabetic	Whether a client is having diabetes or not 0=yes 1=no
region	Whether the client lives in southwest, northwest, southeast or northeast 0=southeast 1=southwest 2=northeast 3=northwest
Charges(Target Variable)	Medical Cost the client pay

Now we examine the other independent variables with the dependent variable (charges).

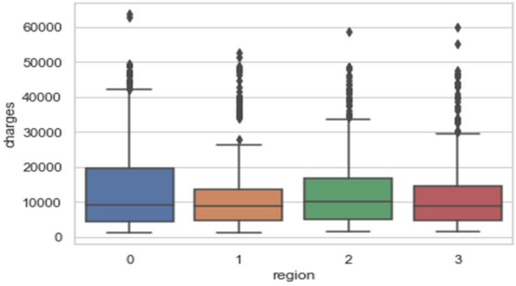


Fig.1.Box plot of Medical Charges per Region

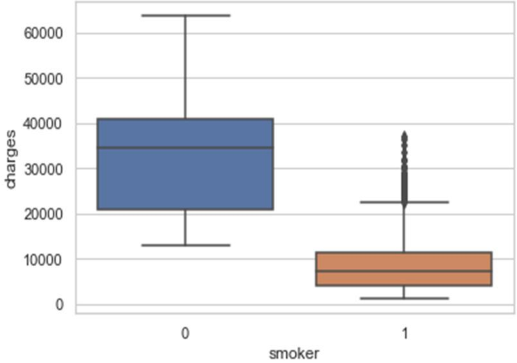


Fig.2.Box plot of Medical Charges by Smoking status

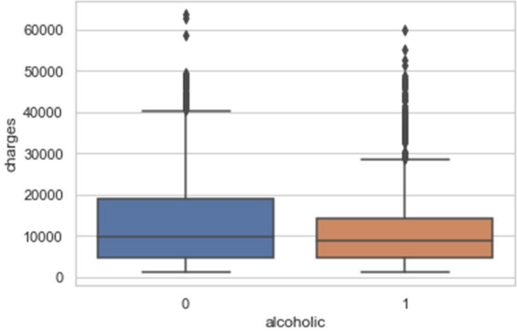


Fig.3.Box plot of Medical Charges for alcoholic person

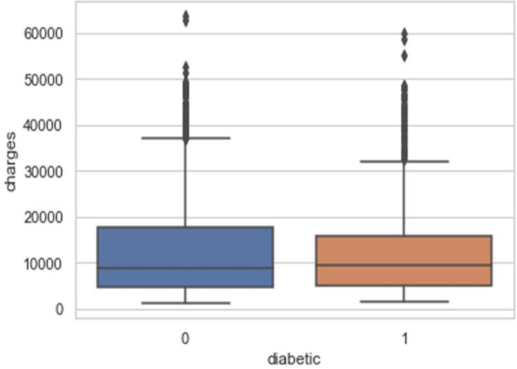


Fig.4.Box plot of Medical Charges by diabetic status

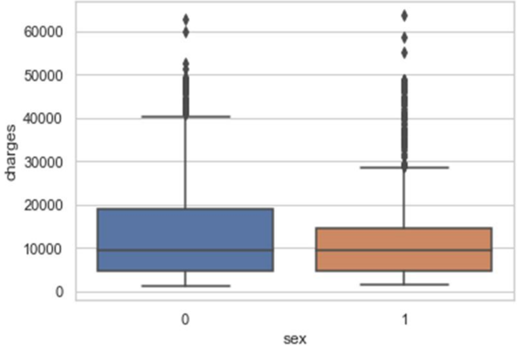


Fig.5.Box plot of Medical Charges per Gender

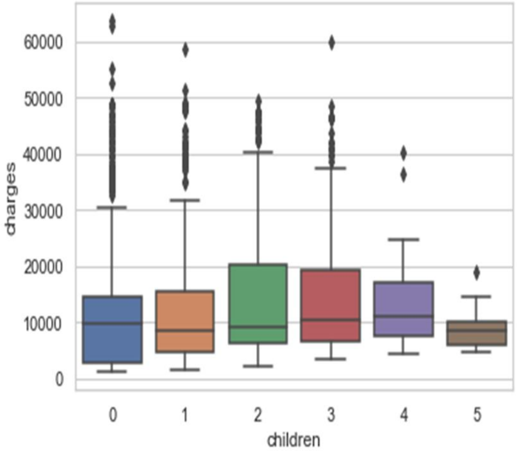


Fig.6.Box plot of Medical Charges per children

Figure 1 shows the impact of the region variable on charges; Figure 2 shows the impact of the smoking variable on charges, and Figure 5 shows the impact of gender variable on charges, and Figure 6 shows the impact of the number of children on charges.
Based on figures 1, 2, 5, and 6, we can say that the region does not have much impact on medical cost. And smokers spend a lot more on medicine. Charges are not affected by Gender but by the number of people. People with two children have more medical expenses.People with five children, however, spend fewer expenses.

IV. RELATED WORKS

A. Literature Review

In this section, analysis efforts from the exploration of knowledge and machine learning techniques are mentioned. Many papers have discussed the difficulty of claim prediction. Jessica suggested, "Predicting motor insurance claims victimization telematics data". This research compared the performance of provision regression and XGBoost techniques to forecast the presence of accident claims by a little range and results showed that as a result of its interpretability and powerful predictability , logistic regression is a better model than XGBoost.
[4]System projected by Ranjodh Singh in 2019, this technique takes photos of the broken automobile as inputs and produces relevant details, akin to prices of repair, to come to a decision on the number of claims and locations of damage. so the anticipated automobile insurance claim wasn't taken into consideration within the gift analysis however was focused on scheming repair costs.
Oskar Sucki 2019, the aim of this analysis is to check the prediction of churn. Random forests were thought-about to be the simplest model (75 % accuracy). In some fields, the information set had missing values. Following associate degree analysis of the distributions, the choice has been taken to substitute the missing variables with extra attributes suggesting that this data doesn't exist. This is often allowable given that the data is totally haphazardly way} lost, so the missing data mechanism by which the suitable approach to processing is set has 1st to be established.
In 2018, Muhammad rFauzan during this paper, the truth of XGBoost is applied to predict statements. Compare the output with the performance of XGBoost, a group of techniques e.g., AdaBoost, Random Forest, Neural Network. XGBoost offers higher Gini structured accuracy. mistreatment publically accessible urban center Seguro to Kaggle datasets. The dataset includes vast quantities of NaN values however this paper manages missing values by medium and median replacement. However, these simple, unprincipled strategies have additionally proved to be biased. They, therefore, target exploring the cubic centimeter methods that are extremely applicable for the issues of many missing values, such as XGboost.
G. Kowshalya, M. Nandhini. in 2018 classifiers are developed during this study to predict and estimate dishonorable claims and a proportion of premiums for the varied customers based mostly upon their personal and monetary data. For classification, the algorithms Random Forest, J48, and Naïve Bayes are chosen. The findings show that Random Forest exceeds the remaining techniques betting on the artificial dataset. This paper thus doesn't cowl claim forecasts, however rather focuses on false claims .
The on top of previous works failed to contemplate each foreseen the value or claim severity, they solely create a classification for the issues of claims (whether or not a claim was filed for that policyholder) during this study we tend to specialize in advanced applied math ways and machine learning algorithms and deep neural network for predict the value of health insurance.

B. Regression

The multivariate analysis may be a prognostic technique that explores the link between a dependent (target) and also the freelance variable(s) (predictor)[5].

This technology is employed to forecast, estimate model time series, and realize the causative impact relationship among the variables. during this analysis, for example, i need to investigate the relationship between insurance price (target variable) and 6 independent variables supported (age, BMI, kid number, individual living area, or sex and whether or not the client is a smoking person).on the idea of a regression.

The multivariate analysis estimates the connection between 2 or a lot of variables, as declared previously. I used completely different regression models to estimate insurance prices on the premise of six freelance variables, and by exploiting this regression, we are able to forecast future health insurance fees supported by current and past data. There are many blessings of using regression analysis as follows:-

1) It demonstrates the essential relationships between the dependent and experimental variables.
2) It shows the result intensity on the variable quantity of many freelance variables.

Analysis of regression conjointly helps one to match the results of measured variables at varied scales, comparable to independent variable and dependent variable effect [6].These blessings permit market researchers, knowledge analysts, and data scientists to get rid of and verify the most effective variety of variables for prophetical models .

V. REGRESSION MODELS

A. Multiple Linear Regression.

Multiple regression may be an applied mathematics technique which will be wont to analyze the link between one variable and a number of other freelance variables. For example, with the information set utilized in this study, we might need to grasp independent variables (8 independent variables), (linearly) involving the dependent variable (charges). is} mentioned because of the multiple simple regression (MLR) model. associate degree MLR model with t\ independent options and Y results can be calculated as within the following equation

In the higher than equation, u is that the residual regression whereas ὰ is the weight of every independent variable or parameter assigned.

B. Decision trees

DTs are straightforward, terribly popular, fast-training, and straightforward to scan models with comparative or different strategies of learning from the data. they're fairly competent however prone to overfitting in their predictions. they'll be strong by their performance .

C. Gradient Boosting Regression

Gradient boosting algorithmic program is one among the foremost powerful algorithms within the field of machine learning. As we all know that the errors in machine learning algorithms are generally classified into 2 classes i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it's accustomed to minimize bias error of the model [7].

Gradient boosting algorithms are often used for predicting not solely continuous target variables as a regression however additionally categorical target variables (as a Classifier). Once it is used as a regressor, the price operates as Mean square. Error (MSE) and when it is used as a classifier then the price operates as Log loss.

VI. IMPLEMENTATION

The objective of the study is to prophetic the insurance cost supported age, BMI, kid number, the region of the person living, sex, and whether or not a shopper is smoking or not, drinks alcohol or not, having diabeties or not . These options contribute to our target variable prediction of insurance costs.

For the measuring of the value of insurance, many regression models are applied during this study. The dataset is split into 2 sections.

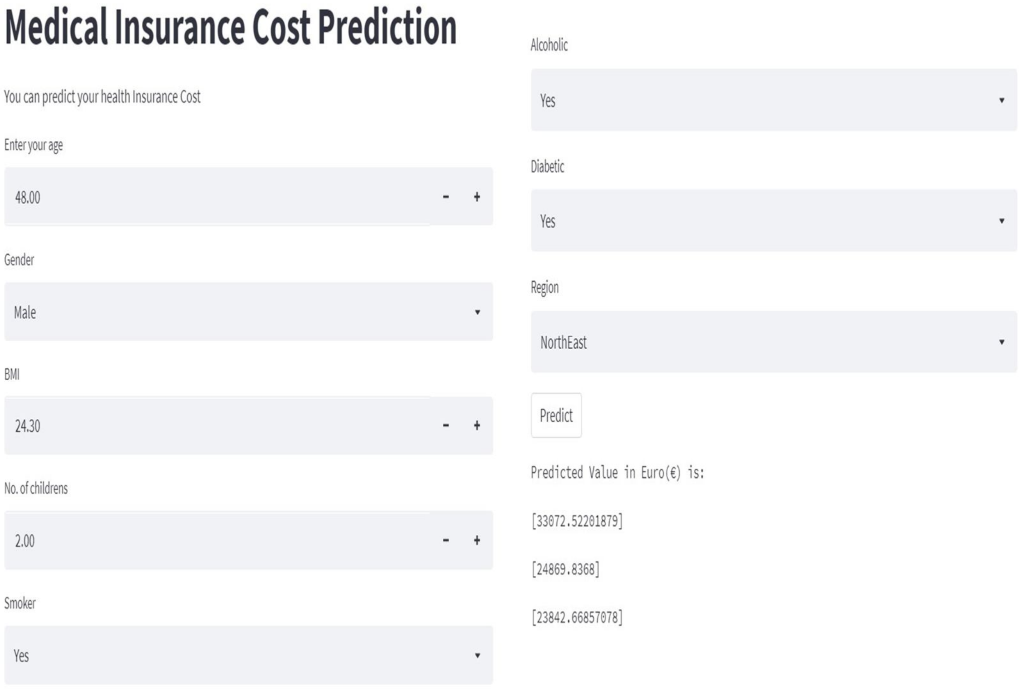
One half for model training and also the other part for model analysis or testing. During this study, the info set is separated into two-part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data. The accuracy is checked with the assistance of r2 score.

Table 3: model performance

Algorithm Used	R2 Score
Linear Regression	74.4738141
Decision Tree Regression	69.0465611
Gradient Boosting Regression	86.8600199

VII. RESULT AND DISCUSSION

OUTPUTS:



Medical Insurance Cost Prediction

You can predict your health Insurance Cost

Enter your age
48.00

Gender
Male

BMI
24.30

No. of childrens
2.00

Smoker
Yes

Alcoholic
Yes

Diabetic
Yes

Region
NorthEast

```
Predicted Value in Euro(€) is:

[33072.52201879]

[24869.8368]

[23842.66857078]
```

VIII. CONCLUSION

The research uses various machine learning regression models to forecast charges of health insurance based on specific attributes, on medical cost personal dataset from Kaggle.com. The findings are summarized in Table 3. shows that Gradient Boosting offers the best efficiency, with an accuracy of 86.86.Gradient boosting can therefore be used in the estimation of insurance costs with better performance than other regression models.

Forecasting insurance prices supported sure factors facilitate insurance suppliers to draw in customers and save time in formulating plans for each individual. Machine learning can considerably minimize these individual efforts in policymaking, as metric capacity unit models can do cost calculation in a very short time, whereas somebody's being would be taking a protracted time to perform constant tasks. This may help businesses improve their profitability. The metric capacity unit models can even manage monumental amounts of data.

REFERENCES

[1] Gupta, S., & Tripathi, P. (2016, February). A leading trend of data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in CS (ICICCS-INBUSH) (pp. 64-69). IEEE.

[2] Yerpude, P., Gudur, V.: Prediction modeling of crime dataset using data mining. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 7(4) (2017)

[3] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Prediction vehicles insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 704. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Vehicle Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[4] Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.

[5] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for not present data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

[6] Grosan, C., Abraham, A.: Intelligent Systems: A Modern Approach, Intelligent Systems Reference Library Series. Springer, Cham (2011)