Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 1816–1822

MEDICAL INSURANCE COST ANALYSIS AND PREDICTION USINGEXTREME GRADIENT BOOSTING ALGORITHM

Dr. S G Balakrishnan [1], Abdulla N [2], Hari Krishnan S [2], Gokul V [2], Amizhthan S P [2]

[1] M.E., Ph.D., Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal
[2] UG Students, Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal

DOI
10.29121/shodhkosh.v5.i6.2024.2595

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

An insurance policy lowers or completely removes the costs related to declining returns caused by different risks. A variety of things affect the cost of insurance. These elements have an impact on how insurance plans are made. In the insurance industry, machine learning (ML) has promise for increasing the effectiveness of insurance policy terms. Actual modelling of insurance claims has emerged as a major field of study in the health insurance industry in recent years, primarily for the purpose of determining appropriate rates. This is essential for drawing in new insured's, keeping the ones you already have, and managing current plan participants well. However, it can be difficult to create an accurate forecast model for medical insurance prices because of the multitude of factors that influence them and their inherent complexities. The expected costs of health insurance could be greatly impacted by a number of factors, such as provider characteristics, lifestyle decisions, health status, accessibility in a given area, and demographic information. Actuarial research into predictive modeling in healthcare is still going strong, as more insurance companies look to leverage ML technologies to increase productivity and efficiency. Regression-based ensemble machine learning models that incorporate different Extreme Gradient Boosting (XGBoost) techniques are used in this study to forecast medical insurance expenses.

Keywords: Medical Insurance, Machine Learning, Model Build, Gradient Boosting, Data Science And Engineering

1. INTRODUCTION

Everyone's health is the centre of their lives. The ability of an individual to cope with their environment on a physical, emotional, mental, and social level is referred to as good health. All people have the best possible lives, unless they have an unforeseen health issue that they cannot predict in advance. Needs such as the desire to possess a home, automobile, or other socially desirable item, as well as other comfortable consumer durables, can be put off if family members have insufficient savings and low-income sources. This isn't the case, though, with the unforeseen medical feature, which requests instant cash, helpful resources, and affects the family's funds. Financial strain due to health issues can undoubtedly destroy a family's long-term financial goals, which can include retirement planning, child marriage or schooling, in addition to the objectives mentioned above. One may also wonder how to overcome a crisis of this nature, and the answer to that is none other than health insurance, which will help safeguard an individual's and their family's well-being without increasing the likelihood of a financial crisis and upsetting financial stability. One type of preferred coverage product is health insurance, which pays for medical and surgical costs for an insured individual, family, or group of individuals. It's an organization where an individual, family, or group acquires health insurance in advance by paying a maximum class fee. Put another way, health insurance is a plan that makes it possible to postpone, reduce, or eliminate costs associated with an insured person's medical bills. In any of the nation's network hospitals, the insurer will either

provide reimbursement for medical expenses spent under the policy or issue certain cashless remedies for medical conditions. There are many threats and uncertainties in the world in which we live. Properties, durables, people, and residences are all vulnerable to different threats, each with a different level of danger. These risks range from the potential for getting sick to passing away without protection to losing money or belongings. The financial industry has developed a number of products that use money to safeguard people and organizations from dangers, even if they may never completely disappear. Therefore, insurance is one of the plans that reduces or completely eliminates loss expenses resulting from a variety of risks. For this specific policy, insurance companies must therefore accurately ascertain the extent of coverage and related insurance expenses. The evaluation of insurance rates is influenced by a wide range of significant elements. When the amounts are computed, the total cost of the policy will vary if any component is added or eliminated. Thus, it is imperative that these activities be completed accurately. Insurance brokers also utilize a range of technology to calculate the insurance premium because human error is a serious risk. ML is therefore helpful in this circumstance. It may be possible to generalize the procedure or work done to create the policy by using machine learning. These machine learning models are self-instructing. The model is trained using historical insurance data. Then, by using the components required to measure the payments as its inputs, the model can reliably forecast the costs of insurance plans. This lowers the requirement for labor and human resources while raising corporate profitability. Therefore, ML can increase accuracy. The data analytics process is shown in fig 1.

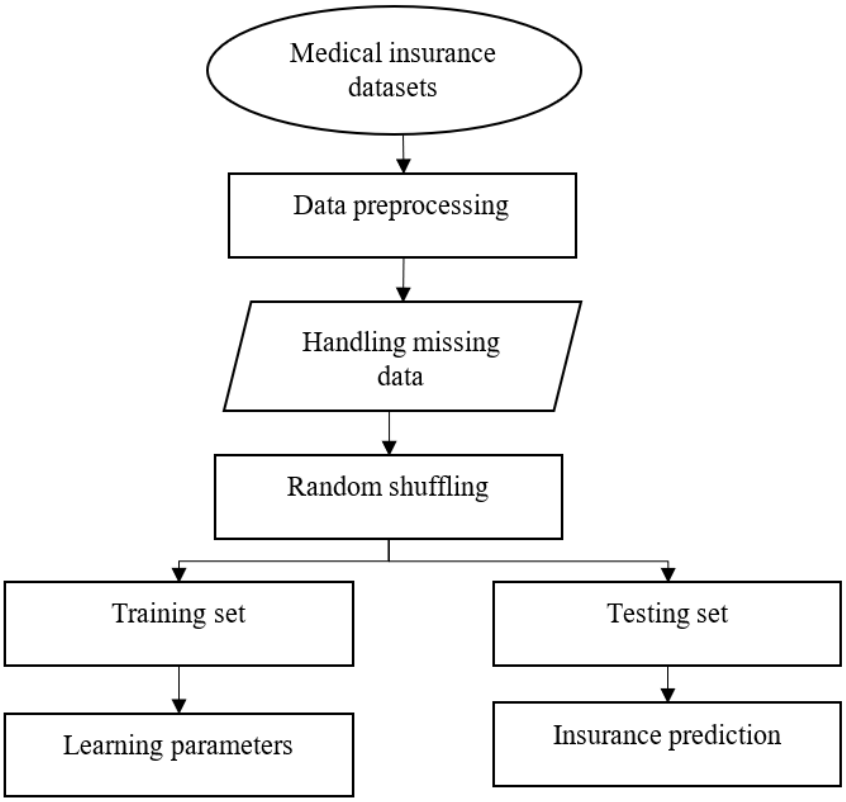

Fig 1: Data analytics process

2. RELATED WORK

Sazzad Hossen, et.al,…[1] used individual and local health data to forecast insurance amounts for various categories of people. The effectiveness of these techniques was investigated using nine regression models: Linear Regression, XGBoost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression. The dataset was used to train the models, and the training data was then used to make some predictions. The model was then put to the test and confirmed by contrasting the actual data with what was predicted to be abundant. Medical insurance is an essential part of the medical sector. However, since patients pay the majority of medical expenses, it is difficult to forecast their expenditure. Data prediction involves a number of machine learning algorithms and deep learning strategies. Evaluation is done on the accuracy and

training time variables. The lot of machine learning algorithms only require a brief time of training. However, the prediction results from theseapproaches are not very accurate. Deep learning models can also find hidden patterns, but their usage in real-time is constrained by the training period

Jonelle Angelo S. Cenita, et.al,…[2] used machine learning algorithms because of their ability to produce results that are effective. Based on a box plot of descriptive statistics and exploratory data analysis, the smoking component is the primary cause of the charge increase. This is because there is a disparity between the two groups, as indicated by the median of one group sitting outside the box of the other group. In light of this, it is highly advised that smoking be eliminated as a major contributor to health insurance costs. A common type of chart used in explanatory data analysis is the Box plot, which makes use of descriptive statistics. A Box plot shows the quartiles (or percentiles) and averages of the data to visually depict the scenes and distribution of the numerical data. Because the boxes and interquartile range in Figure 8's box plot do not overlap, it is evident that there is a difference between the two groups. Furthermore, the median, or middle line, for each group totally rests outside of each box, suggesting that the two groups are likely different. According to the EDA, the smoking feature is what leads to the highest premiums for health insurance.

Chinthala Shreekar, et.al,…[3] Examined a few common regression models for this research. But going forward, there are some new methods that must be handled, like Random Forest Classifier, Decision Tree (CART), XGBoost, Support Vector Machine (SVM), and Stochastic Gradient Boosting. In addition to model evaluation, a variety of optimization techniques can be used, including the Genetic Algorithm and the Gradient Descent Algorithm. To improve accuracy, we can also apply particular feature selection strategies on our dataset before training our model, as some features may be missed while forecasting charges. Numerous studies on the computation of medical expenses have been published in various settings pertaining to health. Machine learning is based on several likely assumptions, but its effectiveness depends on applying a nearly exact technique. Relating to the specified problem area and building, training, and utilizing the model in accordance with the accepted procedures. In order to determine the cost of an item, Moran and colleagues "used thorough linear regression method ICU using patient profile information and DRGs." The length of time spent in the groups (related to the diagnosis) at a hospital, among other factors.

D. Rohan, et.al,…[4]When properly designed, multivariate analysis can be a prognostic technique that examines the relationship between independent variables, also known as predictor variables, and dependent variables, also known as target variables. This method estimates and forecasts the project's time structure to ascertain the association between the variables. For instance, I had to assess the association between six independent factors (age, gender, BMI, number of children, personal region of residence or habitation, and whether or not the client smokes) and insurance rates (the objective variable) throughout this study. Regarding the regression technique. Numerous analyses evaluate the shares of two or more variables, as was previously mentioned. I use a fully variable regression model with six free variables to predict insurance rates. With the help of historical and present data, we can use this regression to predict future health insurance premiums. Regression analysis has the following benefits, among others: - It illustrates the relationship between variables and experimental variables. It demonstrates thatthe strength of the findings is greater than the number of variables for a large number of free variables. By calculating the difference between various indicators in relation to independent and linked variables, regression analysis facilitates result comparison. These findings allow data scientists, information analysts, and business researchers to create and implement the most effective techniques for multiple variable predictions at their own standards.

Mohamed hanafy, et.al,…[5] Using a personal medical cost data set from Kaggle.com, apply deep neural networks and various machine learning regression models to predict health insurance costs based on specific features. Stochastic Gradient Boosting has the maximum efficiency, as shown by Table IV, which summarizes the data, with an accuracy of 85.82, an MAE value of 0.17448, and an RMSE value of 0.380189. Thus, when it comes to estimating insurance costs, stochastic gradient boosting outperforms other regression models. Insurance policy providers can save time and draw in customers by predicting insurance costs based on specific criteria rather of having to create policies for each individual. Machine learning has the potential to significantly minimize these individual efforts in policymaking because ML models can calculate costs fast, whereas a human would need a lot of time to accomplish the same work. Businesses will become more profitable as a result. The ML models can also manage enormous amounts of data.

3. EXISTING METHODOLOGIES

Medical insurance is widely acknowledged as a crucial component of the medical field's infrastructure by experts and professionals. On the other hand, because most medical expenses are covered by individuals with rare illnesses, it is challenging to estimate medical costs. In the process of making predictions, several machine learning techniques are used. However, the accuracy of the results that these approaches predict is not very good. Although machine learning models can find hidden patterns, their training period limits how often they can be used in real time. In order to achieve high forecast accuracy, special ensembles for individual insurance rate estimation are constructed in this study. In this paper, many ensemble models based on boosting, bagging, and assembly processes were used to solve challenges related to medical insurance cost prediction. The experiment's findings demonstrate that, when it comes to finishing the assigned work, the new machine learning-based assembling model performs better than the previous model in terms of prediction accuracy. Most existing prediction models for medical insurance costs follow a predetermined protocol. First, information is gathered from several sources and includes important variables like age, gender, BMI, smoking status, location, past medical history, and pre-existing illnesses. A preprocessing step is started after data collection to clean up the dataset, remove errors, and deal with missing data. Standardizing numerical data and encoding categorical variables for analysis could also be part of this phase. Then, feature selection or engineering methods are used to determine which factors have the greatest influence over insurance premiums. To identify pertinent features, this stage may involve statistical analysis or the use of domain expertise. To improve model performance, engineers may also create new features based on pre-existing ones. After identifying features, the system selects the best algorithms to train predictive models in the model selection step. These algorithms might be more complex machine learning techniques like random forests or decision trees or linear regression.

4. PROPOSED WORK

Almost everyone is connected to a commercial or public health insurance provider these days. One must have health insurance. Each organization's insurance needs are based on a number of criteria. Moreover, rural communities are less likely to be aware of the fact that the Indian government provides free health insurance to anybody who qualifies as impoverished. Due to the complexity of this process, some rural residents opt to obtain private health insurance or make no investment in health insurance at all. Furthermore, people are susceptible to deception regarding the extent of coverage and may overpay for health insurance. The purpose of this paper is to provide readers an idea of how much is needed based on their individual health situation. After that, clients may follow the benefits and policies of any health insurance company while bearing in mind the approximate amount from our project. This can help someone focus less on the unnecessary features of insurance and more on its positive aspects. In this study, we might recommend the Gradient Boosting Algorithm. Gradient boosting is an ensemble strategy that improves performance by combining multiple weak models and building on them. The basic idea behind gradient boosting is to compute a sequence of simple trees, each of which is built using the prediction residuals of the previous one. Through steadily decreasing the errors in each new tree, the method improves the model's predictions. Fig 2 shows the proposed architecture for medical insurance cost prediction

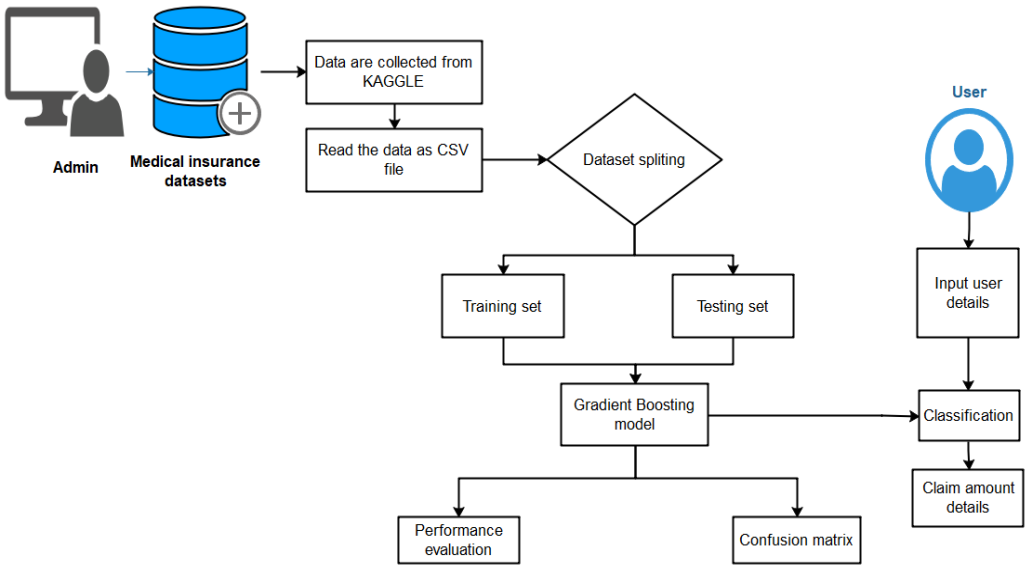

Fig 2: Proposed architecture

Gradient Boosting is a popular algorithm used in machine learning for a variety of tasks, including intrusion detection. The steps involved in using the Gradient Boosting algorithm for intrusion detection are as follows:

Data Collection: Gathering the data needed for the model's training and testing is the initial stage in applying any machine learning technique. When it comes to intrusion detection, this data usually consists of timestamps, packet headers, and payload data from network traffic.

Data Preprocessing: To be used in the machine learning model, the data must be preprocessed after it has been gathered. This process includes feature engineering, feature selection, and data cleaning.

Splitting Data: After preparation, the data is divided into training and testing sets. After the model has been trained using the training set, its performance is evaluated using the testing set.

Training the Model: Using the training data, the Gradient Boosting model must next be trained. Using the residual mistakes from the previous tree as training data, the method will iteratively construct a number of decision trees. Minimizing the loss function, which calculates the difference between expected and actual values, is the aim.

Hyperparameters Tuning: Hyperparameters are the parameters that determine how the Gradient Boosting algorithm behaves during training. The algorithm's performance can be improved by adjusting the values of these hyperparameters. Some of the hyperparameters that can be changed are the number of trees, the maximum depth of each tree, and the learning rate.

Testing the Model: The performance of the model is assessed using the testing data once it has been trained. For intrusion detection, accuracy, precision, recall, and F1-score are common evaluation measures.

Deployment: In order to monitor network traffic and identify intrusions in real-time, the trained model can finally be implemented in a production environment. As fresh data becomes available, the model can be updated over time, and its performance can be continuously observed and assessed.

5. EXPERIMENTAL RESULTS

The system for predicting medical insurance costs can be implemented in Python in this study using real-time datasets gathered from KAGGLE online sources. The system's performance as measured by the accuracy parameter.

To assess the system's performance, a variety of performance metrics, including accuracy, sensitivity, specificity, error rate, and precision, can be calculated.

True positive (TP): quantity of precise positive predictions—true positives.

False positive (FP): quantity of inaccurate positive predictions, or false positives.

True negative (TN): quantity of true negatives—the ideal negative forecast.

False negative (FN): quantity of actual negatives - inaccuracies in negative forecasting.

The ratio of the overall number of perfect predictions to the total quantity of test data is known as accuracy (ACC). Another way to show it is as 1 - ERR. The finest possible accuracy is 1.0, whereas the very worst is 0.0.

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \text{ x } 100$$

ALGORITHM	ACCURACY
SVM ALGORITHM	50%
RANDOM FOREST ALGORITHM	65%
GRADIENT BOOSTING ALGORITHM	95%

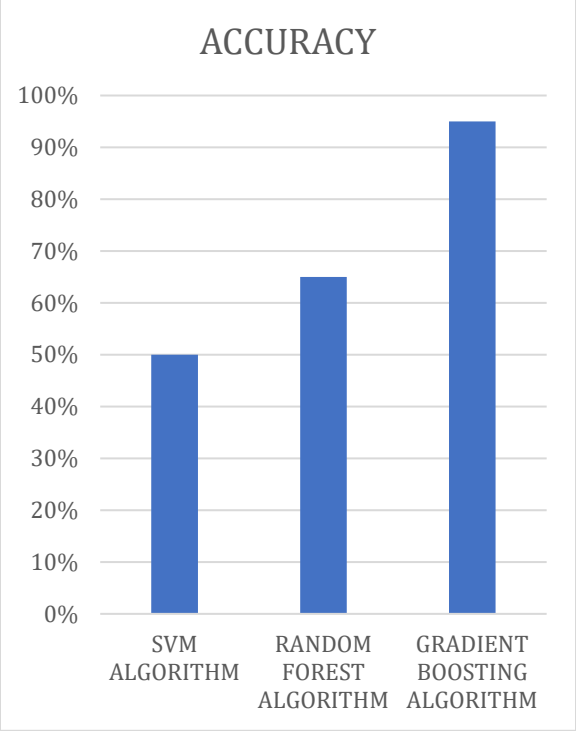

Fig 3: Performance chart

From the above graph, Gradient boosting algorithm provides improved accuracy rate than the existing algorithms

6. CONCLUSION

In conclusion, the utilization of the Gradient Boosting Algorithm for medical insurance cost analysis and prediction yields significant benefits and insights for both insurance providers and policyholders. Through the application of Gradient Boosting, we can effectively handle complex interactions among various factors influencing insurance premiums, such as age, BMI, smoking habits, pre-existing conditions, and geographic location. This algorithm excels in capturing non-linear relationships and feature interactions, allowing for more accurate predictions compared to traditional linear models. By leveraging Gradient Boosting, we have identified key determinants driving medical

insurance costs and developed robust predictive models. These models enable insurance companies to assess risk with greater precision, leading to more tailored pricing strategies that reflect individual healthcare needs and associated expenses. Furthermore, Gradient Boosting enhances the transparency and interpretability of predictive models, providing valuable insights into the relative importance of different factors influencing insurance costs. This transparency fosters trust among stakeholders and facilitates informed decision-making.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Bhatia, Kashish, et al. "Health Insurance Cost Prediction using Machine Learning." 2022 3rd International Conference for Emerging Technology (INCET). IEEE, 2022.

enita, Jonelle Angelo S., Paul Richie F. Asuncion, and Jayson M. Victoriano. "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance." arXiv preprint arXiv:2304.12605 (2023).

hreekar, Chinthala, et al. "Cost Prediction of Health Insurance." International Research Journal of Engineering and Technology 10.01 (2023).

eshav Kaushik,Akashdeep Bhardwaj, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums"International Journal of Environmental Research and Public Health, 2022

ohamed hanafy, Omar M. A. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-10 Issue-3, January 2021

hreekar, Chinthala, et al. "Cost Prediction of Health Insurance." International Research Journal of Engineering and Technology 10.01 (2023).

alyani, G. Satya Mounika. "By contrasting decision trees with logistic regression, a novel categorization-based cost prediction method for health insurance may be developed under supervision." Journal of Survey in Fisheries Sciences 10.1S (2023): 1468-1477.

ijayalakshmi, V., A. Selvakumar, and K. Panimalar. "Implementation of Medical Insurance Price Prediction System using Regression Algorithms." 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2023.

emp, James, et al. "Context discovery and cost prediction for detection of anomalous medical claims, with ontology structure providing domain knowledge." Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023). SCITEPRESS, California, USA, (to appear). Google Scholar Google Scholar Cross Ref Cross Ref. 2023.

Izmie, A. A., et al. "Healthcare Management and Medical Insurance with Predictive Analytics Using Machine Learning." International Research Journal of Innovations in Engineering and Technology 7.10 (2023): 49.

Izmie, A. A., et al. "Healthcare Management and Medical Insurance with Predictive Analytics Using Machine Learning." International Research Journal of Innovations in Engineering and Technology 7.10 (2023): 49.

Cenita, Jonelle Angelo S., Paul Richie F. Asuncion, and Jayson M. Victoriano. "Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance." arXiv preprint arXiv:2304.12605 (2023).

Li, Zhengxiao, Yifan Huang, and Yang Cao. "Analyzing covariate clustering effects in healthcare cost subgroups: insights and applications for prediction." arXiv preprint arXiv:2303.05793 (2023).

Nalluri, Venkateswarlu, et al. "Building prediction models and discovering important factors of health insurance fraud using machine learning methods." Journal of Ambient Intelligence and Humanized Computing 14.7 (2023): 9607-9619.

Lyu, Yuwen, et al. "Prediction of patient choice tendency in medical decision-making based on machine learning algorithm." Frontiers in Public Health 11 (2023): 1087358.