International Journal of Trend in Research and Development, Volume 11(2), ISSN: 2394-9333
www.ijtrd.com

Medical Insurance Premium Prediction

[1]B. Sivaiah, [2]Thaili Bhagya Laxmi, [3]B. Sree Harsha and [4]Yasmeen,
[1,2,3]UG Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana.
[4]Assistant Professor, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana.

***Abstract*—** Medical Insurance Premium Prediction is a regression problem where you try to predict the cost of an insurance premium based on certain features like Body Mass Index, number of dependents, smoking habits. Here, we are using the xgboost-extreme-gradient-boosting algorithm for more accuracy. And motive of this project is to get the information about the insurance that should be paid by the person based on their BMI and smoking so that it gives the perfect idea that should be paid annually by the person.We used the algorithms - Linear Regression , Support Vector Machine , Random Forest , Gradient Boosting and Xg Boost extreme gradient boosting algorithm.In our project we used Xg boost extreme gradient boosting algorithms along with the existing algorithms which is better in the context of accuracy.

Keywords*—Healthcare, Insurance, Regression, Machine Learning, Prediction, Data analysis.*

I. INTRODUCTION

Medical insurance is a crucial aspect of healthcare. In earlier times, people depended on traditional methods, where insurance agents were responsible for estimating the specific amount an individual should contribute in premiums for their medical insurance. But now, individuals can find their total insurance amount online itself

Medical Insurance premiums are dependent on various factors like existing diseases, age, BMI, smoking habits, etc. Our study doesn't offer a precise figure for the amount needed by a particular health insurance provider. However, it does provide a broad understanding of the potential costs that an individual might face. Forecasting medical insurance premiums holds significance for both individuals and insurance companies alike.

People assume medical insurance costs to be high due to which they avoid taking medical insurance for themselves and their family, but it has become a necessity now. This project aims to create an automated system capable of forecasting an individual's future medical insurance expenses.In classification, learning algorithms take the input data and map the output to a discrete output like true or false. In regression, learning algorithms map the input data to continuous output like cost.

The goal of this research is to help individuals understand the amount of money they may need for health insurance rather than the unnecessary ones. In the modern world, it is essential to have health insurance, and most people have a relationship with a public or private health insurance provider. The factors that influence insurance costs vary from company to company. Additionally, some people in rural areas may not be aware that are below the poverty line. However, the process can be complex, and some rural residents either get private health insurance or make no investment at all. Moreover, individuals might be at risk of being deceived into purchasing costly health insurance plans that may not be necessary for their needs. Our study doesn't offer a precise figure needed by any specific health insurance provider, but it does provide a general understanding of the potential cost an individual may bear for their own health insurance.

II. LITERATURE SURVEY

This segment presents exploration and machine learning techniques considering insurance prediction. Several studies have highlighted the challenges associated with predicting insurance outcomes. Jessica proposed a study titled "Predicting insurance claims using telematics data," comparing logistic regression and XGBoost techniques. The research concluded that due to its interpretability and strong predictability, logistic regression outperformed XGBoost, achieving an accuracy of up to 78%.

In 2019, Oskar Sucki aimed to assess prediction accuracy, considering random forests as primary effective model with a 75% accuracy rate. The dataset encountered missing values, addressed by substituting them with additional attributes on the premise that the data was randomly lost.

Muhammad rFauzan, in 2018, applied XGBoost to predict statements, comparing it with AdaBoost, Random Forest, and Neural Network. XGBoost demonstrated superior accuracy despite dealing with a dataset containing substantial NaN values, managed through mean and median replacement, acknowledged for its simplicity.

final value or classification. For decision trees used in regression, the measure of impurity shifts from Gini impurity to using the sum of squared residuals or variance.

Referring to "The Supervised Learning Methods for Predicting Healthcare Costs" by Kensaku Kawamoto, the author identifies five methods for predicting medical insurance costs. The evaluation, conducted on a dataset comprising 90,000 individuals, 6.3 million medical claims, and 1.2 million pharmacy claims, highlighted "gradient boosting" as the best-performing method for low to medium- cost individuals. For high-cost individuals, Artificial Neural Network (ANN) and the Ridge regression model demonstrated the highest performance. The study broadly categorized cost prediction methods into rule-based, statistical, and supervised learning. The author acknowledged limitations, including the use of a specific regional dataset and expressed the intention to explore more advanced supervised learning methods, such as deep learning and structural analysis.

III. METHODOLOGY

A. Linear Regression Algorithm

Linear regression is a machine learning algorithm rooted in concept of “supervised learning.” It is employed to forecast the value of a dependent variable (y) depending on the value of an independent variable (x). Essentially, this implies that linear regression is utilized to ascertain how closely a dependent variable is related to an independent variable, and then make predictions according to that correlation (H. Goldstein, 2012).This is an invaluable asset for data analysis, as it allows analysts to understand complex trends and correlations in data, and to enhance the precision of predictions regarding future outcomes.

Linear Regression Formula

y= mx+b

Here,

m=slope

x=Value of first dataset Y=Value of second dataset

The values of x and y in this scenario represents the training datasets used for illustrating the Linear Regression model.

B. Support Vector Machine Algorithm

SVM, is a commonly employed supervised learning technique for solving classifying and predicting tasks, with a focus regarding classification in machine learning (T. Han, 2020). The objective of Support Vector Machines (SVM) is to identify the optimal line or boundary of decision that can effectively divide a multi-dimensional space into distinct classes. This facilitates the accurate classification of new data points in the future. This optimal decision boundary is known as a hyperplane, which is formed utilizing extreme vectors and points called support vectors. SVM is favored for its capability to effectively classify data and manage high dimensional spaces. This is preliminary estimate. This optimal the decision boundary is known as a hyperplane.

SVM Formula

y=Σi=1 to n(wi.xi)+b

where:

wi is a weights xi is a input data.

b is the bias term.

C. Random Forest

The Random Forest methodology utilizing bootstrapping involves the use of multiple decision tress generated from the data and amalgamated via ensemble learning techniques. This approach often leads to accurate predictions and classifications by taking the average of the outcomes from the randomly selected tress.

Random Forest Formula

y=1/N∑I=1 to N tree i(x)

N represents the quantity of trees in the random forest

D. Decision Tree Regression

Decision trees are a category of supervised ML models employed by the Train Using AutoML tool. They classify or regress data based on true or false answers to specific questions. When visualized, the resulting structure takes the form of a tree, featuring various types of nodes, including the root, internal nodes, and leaf nodes. The root node serves as the initial point in the decision tree, from which branches extend to internal nodes and eventually lead to leaf nodes. The leaf nodes represent the ultimate classification categories or real values within the decision tree. Notably, decision trees are renowned for their simplicity and interpretability, making them easy to understand and explain.

To build a decision tree, begin by designating a feature at the root node. Usually, no single feature can perfectly predict the final classes, leading to what is termed impurity. Methods such as Gini, entropy, and information gain are employed to quantify this impurity and determine how effectively a feature classifies the provided data. The feature with the least impurity is chosen as the node at any given level.

For calculating Gini impurity with numerical entries within a feature, start by sorting the information within ascending order and determining the averages of adjacent values. Next, determine the Gini impurity of each selected average value by arranging the data points according to whetherthe feature values are smaller than or greater than the chosen threshold. Evaluate the effectiveness of this selection by assessing how accurately it classifies the data according to the specified conditions. The Gini impurity is then computed using the equation below, where K represents the quantity of classification categories and p denotes the proportion of instances of those categories.

Determine the weighted average of Gini impurities for the leaves associated with each selected value. Select the value with the minimum impurity for that feature. Iterate this procedure for various features to choose the feature and value that will serve as the node. Iterate this procedure at every node and depth level until all data is classified.

Once the tree is constructed, to predict for a data point, traverse the tree using the conditions at each node to reach the

E. Gradient Boosting Algorithm

Gradient boosting is a highly popular machine learning method for analyzing tabular data sets. It is well-known for its ability to handle missing values, outliers, and large categorical values in the features, in addition to its ability to detect nonlinear relationships between the target and the features. This attribute renders it a potent tool for conducting data analysis and making predictions.

Gradient Boosting is a potent boosting algorithm that amalgamates multiple weak learners into a robust learner. In this method, each new model is trained to reduce the loss function, such as average squared deviation or cross-entropy, of the preceding model using gradient descent. This iterative process enhances the overall predictive capability of the model. During each iteration, the algorithm determines the gradient of the loss function concerning the predictions generated by the current ensemble. Subsequently, a new weak model undergoes training to minimize this gradient, refining the overall predictive accuracy of the ensemble.The forecasts of the new model are integrated into the ensemble, and this procedure is iteratively repeated until a predetermined stopping criterion is fulfilled.

Unlike AdaBoost, Gradient Boosting do not adjust the weights of training instances. Instead, each predictor is trained using the residual errors of the preceding model as labels. A specific technique within Gradient Boosting known as Gradient Boosted Trees utilizes CART (Classification and Regression Trees) as its base learner.

Gradient Boosting Algorithm Formula

Y=∑i=1 to N tree i(x)

N represents the number of trees in the gradient boosting ensemble.

F. XG Boost extreme gradient boosting algorithm

XGBoost, or Extreme Gradient Boosting, is a state-of-the-art machine learning algorithm well known for its exceptional predictive performance. It's the gold standard in ensemble learning, especially when it comes to gradient-boosting algorithms. It develops a series of weak learners one after the other to produce a reliable and accurate predictive model.Fundamentally, XGBoost builds a strong predictive model by combining predictions of several weak learners, usually decision trees. It uses a boosting technique to construct

an extremely accurate ensemble model by having each weak learner after it correct the mistakes of its predecessors.

The optimization method (gradient) minimizes the cost function by repeatedly changing the model's parameters in response to the gradients of the errors. The algorithm also presents the idea of "gradient boosting with decision trees," in Y=∑i=1 to N tree i(x)

N represents the number of trees in the xg-boost ensemble.

which the objective function is reduced by calculating the importance of each decision tree that is added to the ensemble in turn.

adding a regularization term and utilizing a more advanced optimization algorithm, XGBoost goes one step further and improves accuracy and efficiency.

It has gained popularity and widespread usage because it can handle large datasets in a variety of machine-learning tasks, including regression and classification.

IV. EXPERIMENTAL ANALYSIS

A. Input Data Used: We extracted the datasets from google for the purpose of training and testing. This dataset is saved in a CVS file and is well organized. To precisely forecast health insurance costs, it's essential to preprocess the dataset by removing any inconsistencies or irrelevant information before applying regression algorithms. The data shows the age and smoking status have the most noteworthy impact on the amount of insurance, with smoking having the greatest effect. However, factors such as BMI and age also play a vital role. Children's property was discovered to possess little impact on the prediction, so it was removed from the input for the model for regression to improve efficiency and accuracy. It portray the information of age and smoking status have the moist significant impact on The quantity of insurance, with smoking having the greatest effect. This is an initial approximation and is not associated with any particular company. These algorithms are crafted to perform classifications or predictions through statistical methods, unveiling crucial insights in the processes of data mining.

B. Training: After the required data has been formatted and once prepared, the model can commence its training and testing phases. A key focus during the training phase is choosing Selecting the appropriate model for the task at hand. This may involve decisive the optimal modeling strategy or concluding the best parameter values for a particular model. In some instances, this process is referred to as model selection because various models may executed and the one that performs the best, is ultimately chosen, which is formed using support vectors and points known as support vectors,

C. Prediction: The model used for predicting the insurance sum for health was based on the association between specific features and the label. The precision of this prediction was determined by the degree to which the expected value matches the actual amount. To enhance accuracy, the model utilized different features, methods, and variations in the sizes of train-test splits. The study revealed that the indicated quantity of data employed for training significantly influenced accuracy, indicating that a larger training size resulted in improved outcomes. The model utilized multiple algorithms to anticipate the premium amount, demonstrating the impact of each attribute on the outcome.

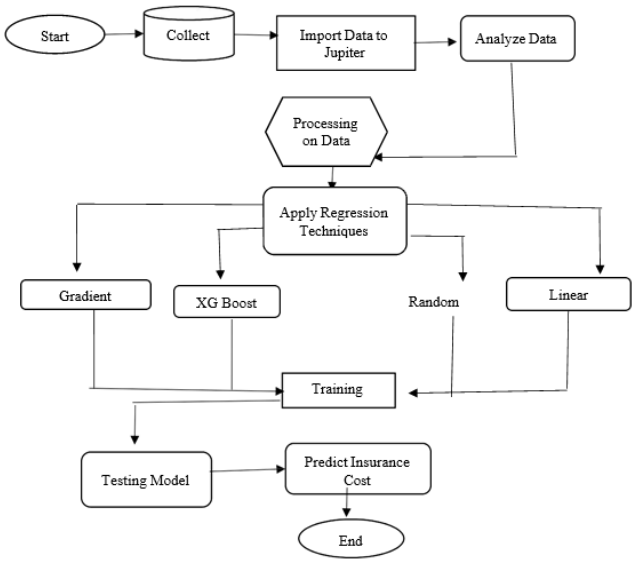

Fig-1 Architecture

Table 1: Dataset Overview

TABLE	DESCRIPTION
Name	Name Of Person
Bmi	Body Mass Index
Age	Age Of Person
Gender	Male/Female
Smoker	Whether A Person Is A Smoker Or Not
Alcoholic	Whether A Person Is Drink Alcohol Or Not
No.Of Kids	Number Of Children The Client Have
Region	Whether The Person Lives In Southwest, Northeast, Southeast Or Northeast
Charges (Target Variable)	Medical Cost The Person Needs To Pay

IV. RESULTS AND DISCUSSION

Algorithms	Accuracy %
Linear Regression algorithm	78.33%
SVM	70.23%
Random Forest	79.00%
Gradient Boosting	81.01%
XG-Boost extreme gradient boosting algorithm	95%

CONCLUSION

In this Project, we have implemented various regression algorithm like linear regression, SVM, Randon Forest, Decision Tres and Xg boost extreme gradient from scratch to predict the medical preces from the input dataset. We also compared the outcomes of Regression Tress, random Forest Regression, Gradient Boosted Regression Tress, and Linear Regression to similar dataset.

While Linear Regression was able to make correct predictions about 78.334% of the time, SVM don't perform well and was not considered a reliable predictor in this case. Xg Boost Extreme Gradient Boosting was determined to be the best model because of its high accuracy rate. The accuracy of this prediction was influenced by how much expected value matched the actual amount.

References

[1] H. Goldstein, W. Browne and J. RasBash, "Multilevel modeling of medical data, "Statistics in Medical, Jhon Wiley and Sons, vol. 21, no 21, pp. 3291-3315, 2012.

[2] Bertsimas, M. V. Bjarnad otter, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, operations Research, vol. 56, no. 6, pp. 1392, 2018.

[3] X. zhu, C. Ying, J. Wang, J. Li, X. Lai et al., "Ensemble of ML-KNN for classification algorithm recommendation," Knowledge-Based System, vol. 106, pp. 933, 2021.

[4] T. Han, A. Siddique, K. Khayat, J. Huang and A. Kumar, "Construction and Building Materials, vol.244, pp.118-271 2020

[5] Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, "Introduction to linear regression analysis", John Wiley & Sons, vol. 821, 2012.

[6] H. Demirtas, "Flexible Imputation of Missing Data", J. Stat. Soft., vol. 85, no. 4, pp. 1–5, Jul. 2018. Available: DOI: 10.18637/ jss. V 085. B 04.

[7] G. Reddy, S. Bhattacharya, S. Ramakrishnan, C. L. Chowdhary, S. Hakak et al., "An ensemble-based machine learning model for diabetic retinopathy classification," in 2020 Int. Conf. on Emergig Trends in Information Technology and Engineering, IC-ETITE, VIT Vellore, IEEE, pp. 1–6, 2020.

[8] Tian Jinyu, Zhao Xin et al., "Apply multiple linear regression model to predict the audit opinion," in 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, IEEE, pp.1–6, 2019.

[9] Ostertagova et al.," Modelling using Polynomial Regression", vol. 48, pp. 500-506, 2012.

[10] Donald W. Marquardt, Ronald D. Snee et al.," Ridge Regression in Practice"," The American Statistician", vol. 29.