Health Insurance Cost Prediction Using Regression Machine Learning Models

Prof. Vijayalakshmi R Y[1], Dr. Manjunatha S[2], Dr. Bharani B R[3], Dr. Preethi S[4]

[1]Assistant Professor, Information Science and Engineering, Cambridge Institute of Technology
[2]Associate Professor, Computer Science and Engineering, Cambridge Institute of Technology
[3]Associate Professor, Information Science and Engineering, Cambridge Institute of Technology
[4]Professor, Information Science and Engineering, Cambridge Institute of Technology

Abstract

India's government spends 1.5 percent of its annual GDP on public healthcare, which is significantly less than that of other countries. Global public health spending, on the other hand, has almost doubled in line with inflation in the last two decades, reaching US$ 8.5 trillion in 2019, or 9.8% of global GDP. Multinational multi-private sectors provide around 60% of comprehensive medical treatments and 70% of out-patient care, which charge patients astronomically high fees. Because of the rising expense of quality healthcare, increased life expectancy, and the epidemiological shift toward non-communicable diseases, health insurance is becoming an essential commodity for everyone. Insurance data has increased dramatically in the last decade, and carriers now have access to it. The health insurance system explores predictive modeling to boost its business operations and services. Computer algorithms and Machine Learning (ML) is used to study and analyze the past insurance data and predict new output values based on trends in customer behavior, insurance policies, and data-driven business decisions, and support in formulating new schemes.

Keywords— Machine Learning, Regression Models, Linear Regression, Support Vector Regression, Random Forest Regression and Gradient Boosting Regression.

1. Introduction

Public health is an integral part of the society, of the country and of the world we live in and is an important matter of concern. Human lives and public health may be endangered by natural calamities, global epidemic and pandemics, global crisis of medical aids, etc. which increases the vulnerability of public health. People can meet with unavoidable and unforeseen circumstances at any point of their lifetime. Individuals, families, companies and properties are uncovered and uninsured to diverse hazard forms, natural calamities, and the likelihood can shift. These perils include the possibility of mortality, health, and property disaster or resource depletion.

People's lives revolve around two main elements: life and prosperity. However, keeping a safe and sound distance from unforeseen events is impossible. The Government of India spends 1.5% of GDP on public health, the lowest level globally [1]. In the last two decades, universal health care has been almost doubled, surpassing US $ 8.5 trillion in 2019, or 9.8% of global GDP [2]. The multi-private sector having state-of-the-art facilities provides almost 60% of total hospitalizations and 70% outpatient services [3]. Thus, Health insurance is becoming an essential commodity for every individual because of the increasing cost of quality healthcare combined with higher life expectancy and widespread transition towards non-communicable diseases. To alleviate these types of problems, the world of fund has developed a variety of tools to protect people and organizations from such unseen catastrophic situations, through the utilization of monetary capital to repay and compensate them. In this manner, insurance is an arrangement that diminishes or evacuates misfortune costs brought about by different dangers.

Today, data has evolved drastically since the recent past decade and insurance carriers have access to it. The health insurance system is exploring ways to use predictive modeling to boost their business operations and services. Computer algorithms and Machine Learning (ML) are used to study and analyze the historical insurance data and predict new output values based on trends in customer behavior, insurance policies and data-driven business decisions, and supports in formulation of new schemes. Besides, most insurance companies use conventional databases to store their data which is primarily structured data. Moreover, merely 10- 15 percent of the total data available is processed for gaining insights. Thus, transformation of the data is necessary to gain valuable insights that may be very crucial for the growth of such companies. Therefore, analyzing the structured data as unstructured data for making better decisions requires statistical and Machine Learning (ML) techniques. The main advantage of ML is that it can be effectively applied to a massive volume of structured, semi-structured, or unstructured datasets. The ML model can be used across multiple value chains to understand the weight-age of risk involved, claims made and customer behavior with greater predictive accuracy. ML applications in the health insurance sector include various tasks such as understanding risk tolerance and premium leakage leading to inaccurately pricing of premiums, loss deterrence, claims handling, expense management, subrogation, litigation, and fraud identification.

When it comes to the value of health insurance in people's lives, it's vital for insurance firms to be as explicit as possible for measuring the sum secured by this approach and the protection charges which must be paid for it. The calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. The intervention of humans in this process may sometime produce faulty or inaccurate results. Additionally, as the data increases manual calculations becomes lethargic and time consuming. Again, in such scenarios the implementation of ML models can be very beneficial for such companies. Therefore, ML may generalize the exertion or strategy to define such an approach. These models can perform self-learning to predict the cost of insurance using past insurance data of the companies. The model inputs are the main parameters that are utilized to calculate the instalments made. This enables the algorithm to precisely estimate the disbursement of insurance coverage. In this way, the correctness can be progressed with ML. The objective of the proposed model is to perform rapid estimation and prediction of insurance charges at a hospital incurred by a patient, using ML models upon the Kaggle dataset. Thus, this paper develops a real-time insurance cost price prediction system named ML Health Insurance Prediction System (MLHIPS) using ML algorithms which will aid the insurance companies in the market for easy and rapid determination of values of premiums and thereby curb down health expenditure. The proposed model incorporates and demonstrates different models of regression such as Support Vector Regression (SVM), Random Forest Regression, Linear Regression and Gradient Boosting Regression to anticipate insurance costs and assess model outcomes. In the proposed model, the Gradient Boosting Regression model has achieved better results with 0.87 compared to all the other models.

2. Related Work

Several studies on estimating medical prices have been published in the health field in different contexts [4], [5], [6], [7], [8]. Machine learning has many probable assumptions, but its performance relies on picking a nearly precise algorithm for the specified problem domain and following the appropriate procedures to build, train, and deploy the model.

Moran et al. [4] “utilized a comprehensive linear regression method which was used to predict the cost of an intensive care unit (ICU) using patient profile data, DRGs (diagnosis-related groups), the length of time spent in the hospital, and additional traits as features.”

Sushmita et al. [5] “proposed a model based on the medical and past expense history of a person to predict his/her future medical costs. Quarterly projected expenses for the future 3,6,9 and 12 months were estimated with the use of the model. They have used random forest and linear regression analysis to predict the costs.”

Lahiri et al. [6] “used a classification algorithm to predict whether an individual's medical expenses would increase in the next year, taking into account the medical expenses of the previous year.”

Gregori et al. [7] have used the logit model and the OLS method to study the multivariate modeling of healthcare costs data.

Bertsimas et al. [8] use data mining techniques, explicitly clustering algorithms and classification trees, and insurance claim data of nearly 500,000 members throughout a three-year period. Based on the data gathered from the medical expenses from the first two years, a justified third-year health-care cost projection is made.

3. Methodology

The regression techniques used are the statistical method that establishes the association between a target or dependent variable and a set of independent or predictor variables. It assumes that both the target and the predictor variables are having numerical values and there exists some kind of correlation between the two. The models that we are implementing in our problem are discussed below,

A. Model Selection:

1. Simple Linear Regression: In simple linear regression [16], the target variable(Y) is dependent on a single independent variable(X) and the model establishes a linear relationship among these two variables.

The linear regression model tries to fit the regressor

line between the independent(X) and dependent(y) variable. The equation of the line is given by:

$$Y = a + Bx \ldots\ldots\ldots\ldots\ldots (1)$$

where “a” and “b” are the model’s parameters called as regression coefficients, “a” is the value of the Y intercept that the line makes when X is equal to zero and “b” is the slope that signifies the change of Y with the change of X. More the value of “b” means a small change in X causes a significant change in Y, and vice versa. The value of “a” and “b” can be found by Ordinary Least Square method.

In linear regression models the values predicted may not be accurate always, there will always be some difference, hence we add an error term ϵ to the original equation (1) that accounts for the difference and thus help in making better predictions.

$$Y = a + bX + \epsilon \text{ .... } (2)$$

Assumptions in Linear Regression

- The sample size of data should exceed the number of available parameters.
- Only over a restricted range of data the regression can be valid.
- Error term is normally distributed. This also means that the mean of the error has expected value of 0.

2. Multiple Linear Regression: Similar to simple linear regression, multiple regression [17] is a statistical procedure that examines the degree of association between a set of independent variables and a dependent variable. There is just one independent and one dependent variable in basic linear regression, but there are numerous predictor variables in multiple linear regression and the value of dependent variable(Y) is now calculated depending on the values of the predictor variables. it is assumed that there is no dependency among the predictor variables. Suppose if the target value is dependent on “n” independent variables then the regressor fits the regression line in a N dimensional space. The regressor line equation is now modified into:

$$Y = a + b1\text{x1} + b2\text{x2} + b3\text{x3} + \text{..........} + b\text{nxn} + \in$$

Where ‘a’ is the Y – Intercept value and $< b1, b2, b3, \ldots\ldots\ldots$

$, bn >$ are the regression coefficient associated with the n independent variables and $\in$ is the error term.

3. Random Forest Regression: Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn’t depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation.

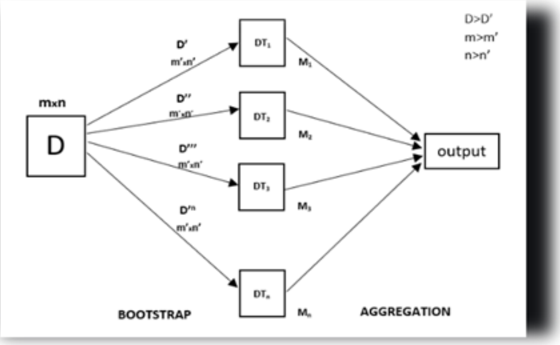


Figure 1

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models.

When the slope (b) of the line is steep then the target variable(Y) is very sensitive to relatively small change in the predictor variable variable(X). In ridge regression by the addition of the lambda value the sensitivity decreases. If lambda is zero then ridge regression reduces to linear regression and when lambda increases gradually the slope the line decreases asymptotically. To know which value of lambda is to choose we try different values of lambda and use cross validation to determine which one result in the lowest variance.

4. Support Vector Regression: In machine learning, Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. In Support Vector Regression, the straight line that is required to fit the data is referred to as hyperplane. The objective of a support vector machine algorithm is to find a hyperplane in an n-dimensional space that distinctly classifies the data points. The data points on either side of the hyperplane that are closest to the hyperplane are called Support Vectors. These influence the position and orientation of the hyperplane and thus help build the SVM.

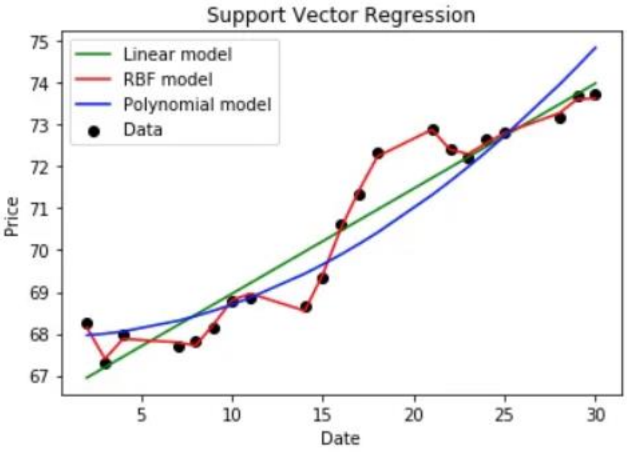
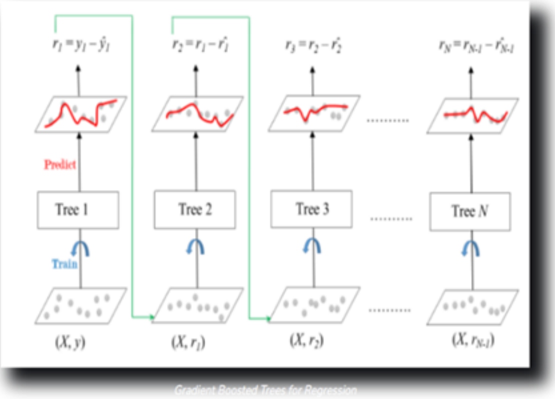


Figure 2

5. Gradient Boosting Regression: Gradient Boosting Regression is a popular machine learning technique for regression problems. It is an ensemble learning method that combines multiple weak learners to create a strong predictive model. The key idea behind gradient boosting is to iteratively add new weak models to the ensemble, each one trained to correct the errors of the previous models. In gradient boosting regression, the weak learners are usually decision trees, and the objective is to minimize the mean squared error (MSE) between the predictions and the actual values. The process of adding new trees to the ensemble is guided by the gradient of the MSE with respect to the predicted values. The gradient indicates the direction of the steepest descent in the error surface, and the new tree is trained to minimize the error along that direction. The most popular implementation of gradient boosting regression is the XGBoost algorithm, which is known for its efficiency and scalability. XGBoost uses a number of optimization techniques to speed up the training process and reduce the memory footprint of the model. It also includes several regularization techniques to prevent overfitting and improve generalization performance.



B. Description of the Dataset:

From the Kaggle site [21] we obtained our dataset for developing the ML Health Insurance Prediction System (MLHIPS). The data set obtained contains seven attributes or features and 1338 rows; out of the seven attributes or features three of them contains categorical values and the rest contain numerical values. The data set is then divided into two halves. The first component is referred to as training data, while the second is referred to as test data. The more data that is supplied to the model during its training period, the more accurate the model will be when making predictions on unseen data. Data is typically split at a ratio of 80:20 for testing and training purposes. Training datasets are used to build models as predictors of health insurance costs, and test sets are used to evaluate regression models. Table.1 displays the dataset's description.

The dataset contained missing values in certain fields. After reviewing the distributions, it was decided to replace the missing variables with new attributes, implying that the data is missing. [9]. This is only possible if the data is lost completely at random; thus, the missing data mechanism, which determines the best method to data processing, must first be developed. [10] [11]. Multilevel structure and hidden dependencies are present in medical data [12]. It is vital to figure out these hidden patterns and use various fundamental analysis techniques present in a combined fashion. This is the reason why in the field of medical data analysis, many researchers use different ensemble Machine Learning models. In order to tackle the problem of price prediction, researchers have also used hierarchical regression analysis. Many of them have used different ensemble learning techniques such as Random Forests, Adaboost, GBM, and XGBM. In paper [13], to forecast the modulus of elasticity of the collective recycled concrete, an ensemble of Gradient Booster, Random Forests (RF) and Support Vector Machines (SVM) is utilised. This classical ensemble has a much higher level of precision.

C. Data Pre-processing:

The dataset contains seven variables, as shown in the table above. While calculating the cost of the Charges of a customer which is our target variable the values of the rest six of the variables are taken into consideration. In this phase, the data is reviewed, properly reconstructed, and properly applied to machine learning algorithms. The dataset was first checked for missing values. The dataset was found containing missing values in the bmi and charges columns. The missing values were imputed by the mean values of the respective attribute values.

As regression models accept only numerical data, the categorical columns in our case the sex, smoker and region columns containing categorical columns were converted into numerical values using label encoding. Then the updated dataset was partitioned into training and testing dataset. And the model was trained using the training dataset.

TABLE I DATASET

Name	**Description**
Age	Customer's Age
BMI	Body mass index of the customer
Number of kids	Number of kids of the customer
Gender	Male / Female
Smoker	Whether the customer is smoker or not.
Region	Where the customer lives: southwest, southeast, northeast, northwest
Charges (target variable)	Medical fee the customer has to pay

RESULTS

The regression model's performance is evaluated on the basis of the following metrices

R^2_Score

Root Mean Square Error (RMSE)

R^2_Score: R-Squared is a good measure to evaluate the model fitness. The R-squared value lies between 0 to 1 (0% to 100%). Large value represents a better fit.

$$R2 = 1 - SSE$$

[redacted] T

where SSE (Squared sum of error): sum of the squared residuals, which is squared differences of each observation from the predicted value. $\sum(yi - y*)^2$ and,

SST (Sum of Squared Total): squared differences of each observation from the overall mean. $\sum (yi - y\wedge)^2$, where

y^i represents the observed value, $y*$ is the projected value and

$y\wedge$ is the average of the observed values.

RMSE: The Root Mean Square error is a common method of calculating a model's prediction error. The RMSE, which represents how close the observed data points are to the model's predicted values, shows the model's absolute fit to the data points. A better match is indicated by values lower RMSE

$$RMSE = \frac{\sqrt{\sum_{i=1} n(\underline{yi- y*})^2}}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

Table Ii. Calculated Results

Regressor	R2_Score	RMSE	Accuracy (in %)
Linear Regression	0.78	4186.50	78.33
Support Vector Regression	-0.07	8592.42	-7.22
Random Forest Regression	0.86	2478.34	86.25
Gradient Boosting Regression	0.87	2447.95	87.79

The outcomes of the above discussed models after testing the models on the test dataset were noted. Table.2 displays the R2, RMSE and Accuracy metrics of the models discussed so far.

Conclusion

Calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. Following human intervention in the process may sometime produce faulty or inaccurate results and also when the data increases the time taken for calculation by human's increases. In scenarios like these the implementation of Machine Learning models can be very beneficial to the company. In this paper, several Machine Learning regression models are used to predict the cost of health insurance based on specific attribute values present in the dataset. The results obtained are summarized in Table II. With an RMSE of 2447.95, an R2 of 0.87, and an accuracy of 87.79 percent, Gradient Boosting Regression is the most efficient. Based on the model's configuration parameters which are tuned during the training phase, on the basis of performance the different proposed models are arranged accordingly starting from Gradient Boosting Regression, Random Forest Regression, Support Vector Regression, and Linear Regression. These models can be incorporated by the companies for calculating the charges in a fast and reliant manner thereby saving the time and cost of the company. These models in the later stage of life cycle can be deployed onto the cloud platforms when the data increases integrated with high end computing resources for faster processing of real time data in short span of time interval.

References

1. "Global Expenditure on Health", WHO annual report 2021, [Online].Available:https://www.who.int/newsroom/events/detail/2021/1 2/15/default-calendar/global-spending-on-health-2021
2. "Health Insurance of India's missing middle", Niti Ayog India, Oct 2021, [Online]. Available: https://www.niti.gov.in
3. "National Health Accounts," National Health Systems Resource Centre. [Online].Available:https://nhsrcindia.org/national-health-accounts records
4. 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