Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Health Insurance Amount Prediction

Nidhi Bhardwaj , Rishabh Anand
Delhi, India
Dr. Akhilesh Das Gupta Institute of Technology & Management

***Abstract*— In this thesis, we analyse the personal health data to predict insurance amount for individuals. Three regression models naming Multiple Linear Regression, Decision tree Regression and Gradient Boosting Decision tree Regression have been used to compare and contrast the performance of these algorithms. Dataset was used for training the models and that training helped to come up with some predictions. Then the predicted amount was compared with the actual data to test and verify the model. Later the accuracies of these models were compared. It was gathered that multiple linear regression and gradient boosting algorithms performed better than the linear regression and decision tree. Gradient boosting is best suited in this case because it takes much less computational time to achieve the same performance metric, though its performance is comparable to multiple regression.**

***Keywords*— Regression, Premium, Machine Learning.**

I. INTRODUCTION

The goal of this project is to allows a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount

needed. Where a person can ensure that the amount he/she is going to opt is justified. Also it can provide an idea about gaining extra benefits from the health insurance.

II. DATASET USED

The primary source of data for this project was from Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow ‘age’, ‘gender’, ‘bmi’, ‘children’, ‘smoker’ and ‘charges’ as shown in Fig. 1. The data was in structured format and was stores in a csv file.

Dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

In a dataset not every attribute has an impact on the prediction. Whereas some attributes even decline the accuracy, so it becomes necessary to remove these attributes from the features of the code. Removing such attributes not only help in improving accuracy but also the overall performance and speed.

	age	gender	bmi	children	smoker	charges
0	19	female	27.900	0	yes	16884.92400
1	18	male	33.770	1	no	1725.55230
2	28	male	33.000	3	no	4449.46200
3	33	male	22.705	0	no	21984.47061
4	32	male	28.880	0	no	3866.85520

Figure 1: Sample of Health Insurance Dataset

In health insurance many factors such as pre-existing body condition, family medical history, Body Mass Index (BMI), marital status, location, past insurances etc affects the amount. According to our dataset, age and smoking status has the maximum impact on the amount prediction with smoker being the one attribute with maximum effect. Children attribute had almost no effect on the prediction, therefore this attribute was removed from the input to the regression model to support better computation in less time.

III. MACHINE LEARNING

Machine learning can be defined as the process of teaching a computer system which allows it to make accurate predictions after the data is fed.

However, training has to be done first with the data associated. By filtering and various machine learning models accuracy can be improved. Fig. 2 shows various machine learning types along with their properties.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

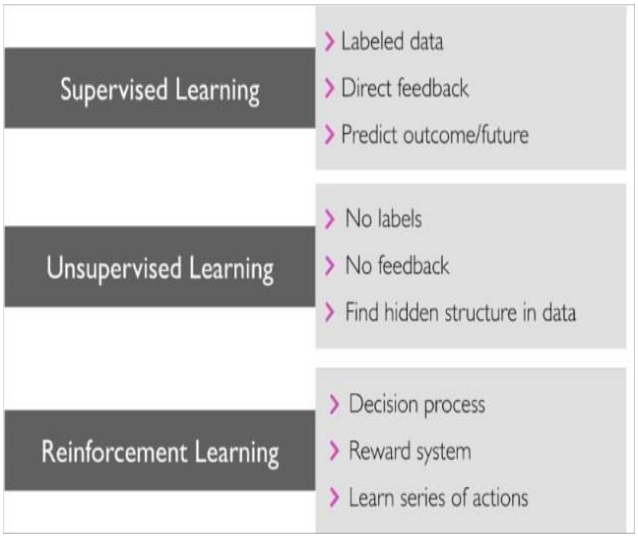

Figure 2: Types of Machine Learning

A. *Supervised Learning*

Supervised learning algorithms create a mathematical model according to a set of data that contains both the inputs and the desired outputs. Usually a random part of data is selected from the complete dataset known as training data, or in other words a set of training examples. Training data has one or more inputs and a desired output, called as a supervisory signal. What's happening in the mathematical model is each training dataset is represented by an array or vector, known as a feature vector. A matrix is used for the representation of training data. Supervised learning algorithms learn from a model containing function that can be used to predict the output from the new inputs through iterative optimization of an objective function. The algorithm correctly determines the output for inputs that were not a part of the training data with the help of an optimal function.

B. *Unsupervised Learning*

In this learning, algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. Test data that has not been labeled, classified or categorized helps the algorithm to learn from it. What actually happens is unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. The main application of unsupervised learning is density estimation in statistics. Though unsupervised learning, encompasses other domains involving summarizing and explaining data features also.

C. *Reinforcement Learning*

Reinforcement learning is class of machine learning which is concerned with how software agents ought to make actions in an environment. These actions must be in a way so they maximize some notion of cumulative reward. Reinforcement learning is getting very common in nowadays, therefore this field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulated-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms.

IV. REGRESSION

Regression analysis allows us to quantify the relationship between outcome and associated variables. Many techniques for performing statistical predictions have been developed, but, in this project, three models - Multiple Linear Regression (MLR), Decision tree regression and Gradient Boosting Regression were tested and compared.

A. *Multiple Linear Regression*

Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable) and the variables being used in predict of the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

B. *Decision tree regression*

Regression or classification models in decision tree regression builds in the form of a tree structure. The dataset is divided or segmented into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision tree with decision nodes and leaf nodes is obtained as a final result. These decision nodes have two or more branches, each representing values for the attribute tested. Decision on the numerical target is represented by leaf node. The topmost decision node corresponds to the best predictor in the tree called root node. Numerical data along with categorical data can be handled by decision tress.

C. *Gradient Boosting Regression*

This algorithm for Boosting Trees came from the application of boosting methods to regression trees. The basic idea behind this is to compute a sequence of simple trees, where each successive tree is built for the prediction residuals of the preceding tree. For predictive models, gradient boosting is considered as one of the most powerful techniques.
Gradient boosting involves three elements:

1. An optimized loss function.
2. An additive model to add weak learners to minimize the loss function.
3. A weak learner to make predictions

V. DESIGNING AND IMPLEMENTATION

A. *Data Preparation & Cleaning*

The data has been imported from kaggle website. The website provides with a variety of data and the data used for the project is an insurance amount data. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

for the project. The data was in structured format and was stores in a csv file format. The data was imported using pandas library.
The presence of missing, incomplete, or corrupted data leads to wrong results while performing any functions such as count, average, mean etc. These inconsistencies must be removed before doing any analysis on data. The data included some ambiguous values which were needed to be removed.

B. Training

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed.
During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. In fact, the term model selection often refers to both of these processes, as, in many cases, various models were tried first and best performing model (with the best performing parameter settings for each model) was selected.

C. Prediction

The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount.
Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy.
The model predicts the premium amount using multiple algorithms and shows the effect of each attribute on the predicted value.

VI. RESULT

We see that the accuracy of predicted amount was seen best i.e. 99.5% in gradient boosting decision tree regression. Other two regression models also gave good accuracies about 80%
In their prediction. Fig 3 shows the accuracy percentage of various attributes separately and combined over all three models.
Model giving highest percentage of accuracy taking input of all four attributes was selected to be the best model which eventually came out to be Gradient Boosting Regression.

	Linear Regression	Decision Tree Regression	Gradient Boosting Regressor
Age	8-13	2-13	2-20
Gender	0	0	0
Smoker	50-64	57-69	57-70
BMI	0-4	0-1	0
Age + Gender	0-15	0-1	0-15
Age + Smoker	9-12	2-4	6-17
Age + BMI	0-9	0	0-11
Gender + Smoker	59-65	56-69	59-67
Gender + BMI	2-10	0	0-3
Smoker + BMI	61-80	58-74	74-81
Age + Gender + Smoker	67-75	57-64	67-75
Age + Gender + BMI	2-17	0	0-14
Gender + Smoke + BMI	59-70	54-77	70-86
Age + Smoke + BMI	67-82	70-79	80-93
Age + Gender + Smoke + BMI	69-82	67-84	90-99.5

Figure 3: Accuracy in percentage (%)

Fig. 4 shows the graphs of every single attribute taken as input to the gradient boosting regression model.

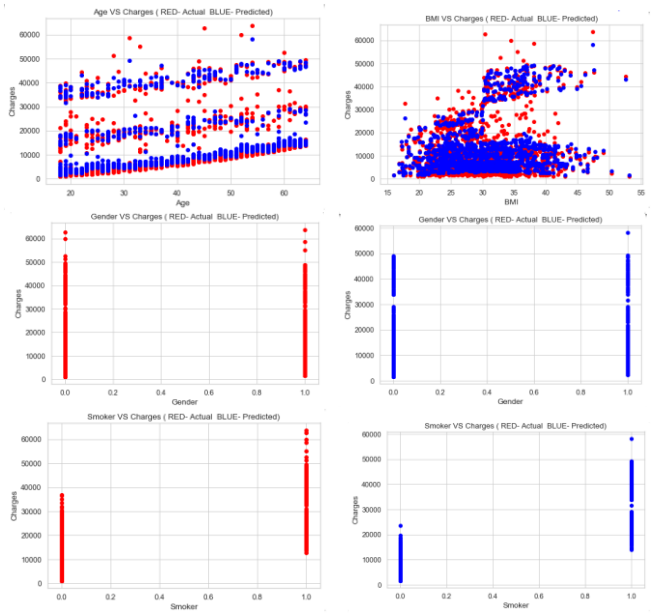

Figure 4: Attributes vs Prediction Graphs - Gradient Boosting Regression

VII. CONCLUSION & FUTURE SCOPE

Backgroun In this project, three regression models are 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 actual premiums to compare the accuracies of these models. It has been found that Gradient Boosting Regression model which is built upon decision tree is the best performing model.

Various factors were used and their effect on predicted amount was examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features.

The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results.

Premium amount prediction focuses on person's own health rather than other company's insurance terms and conditions.
The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.

REFERENCES

[1] https://www.moneycrashers.com/factors-health-insurance-premium-costs/
[2] https://en.wikipedia.org/wiki/Healthcare_in_India
[3] https://www.kaggle.com/mirichoi0218/insurance
[4] https://economictimes.indiatimes.com/wealth/insure/what-you-need-to-know-before-buying-health-insurance/articleshow/47983447.cms?from=mdr

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

[5] https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php
[6] https://www.zdnet.com/article/the-true-costs-and-roi-of-implementing-ai-in-the-enterprise/
[7] https://www.saedsayad.com/decision_tree_reg.htm
[8] http://www.statsoft.com/Textbook/Boosting-Trees-Regression-Classification

.