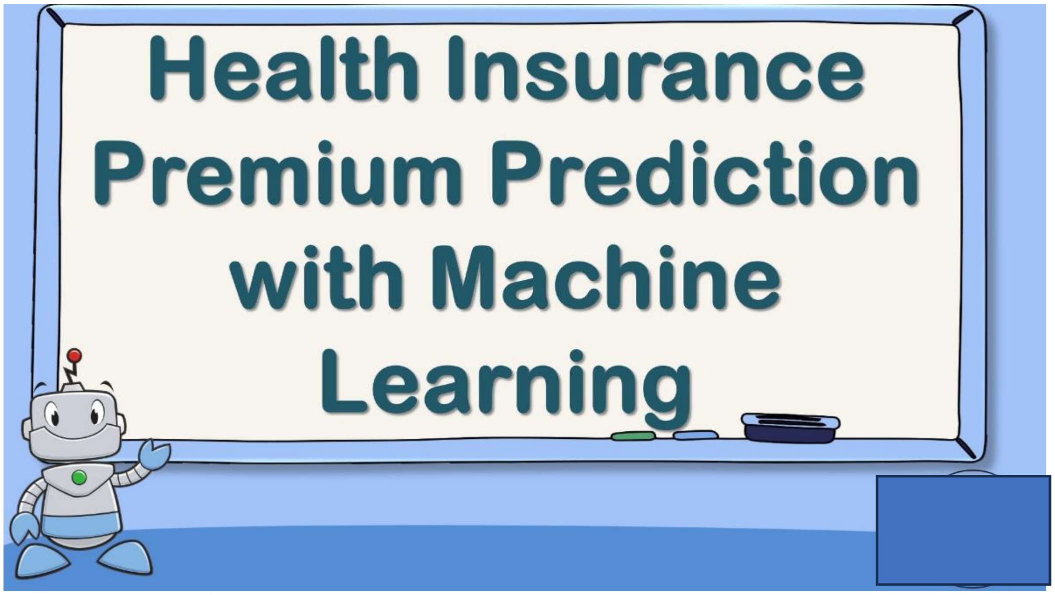


Health Insurance Premium Prediction with Machine Learning

Major Project Health Insurance Cost Prediction

PARTICIPANTS:

- Jagatheesvaran S
- Tharun aaditya
- Yashika Aggarwal
- MD afroz alam
- Prathyanga

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 Insurance rather than the futile part.

Health Insurance is a necessity nowadays, and almost every individual is linked with a government orprivate 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.

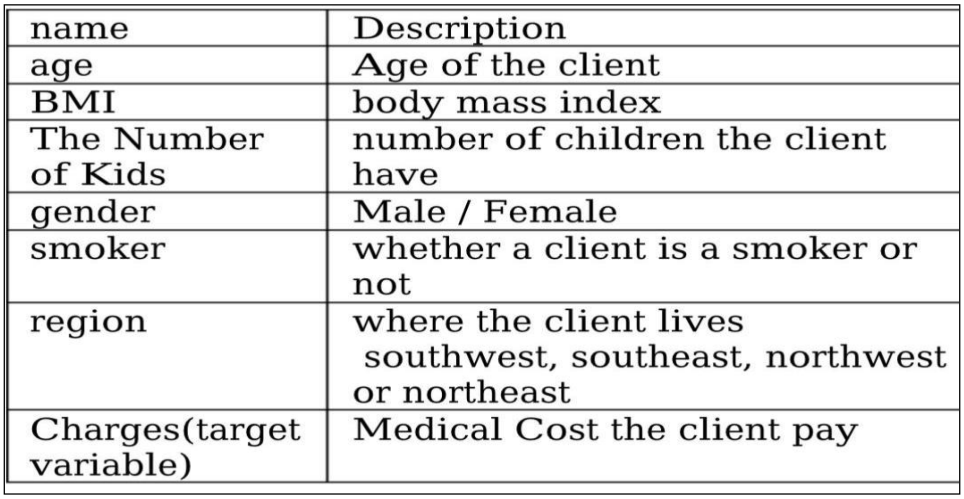
Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount he/she is going toopt is justified. Also it can provide an idea about gaining extra benefits from health insurance.

DATASET

To create the claim cost model predictor, we obtained the data set. The data set includes seven attributes see

Table 1; the data set is separated into two-part the first part is called training data, and the second calledtest data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model. the following table shows the Description of the Dataset.

DATA SET OVERVIEW



name	Description
age	Age of the client
BMI	body mass index
The Number of Kids	number of children the client have
gender	Male / Female
smoker	whether a client is a smoker or not
region	where the client lives southwest, southeast, northwest or northeast
Charges(target variable)	Medical Cost the client pay

DATA SET USED

The primary source of data for this project was Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follows ‘age’, ‘gender’, ‘BMI’, ‘children’, ‘smoker’, and ‘charges. The data was in a structured format and was stored in a CSV file.

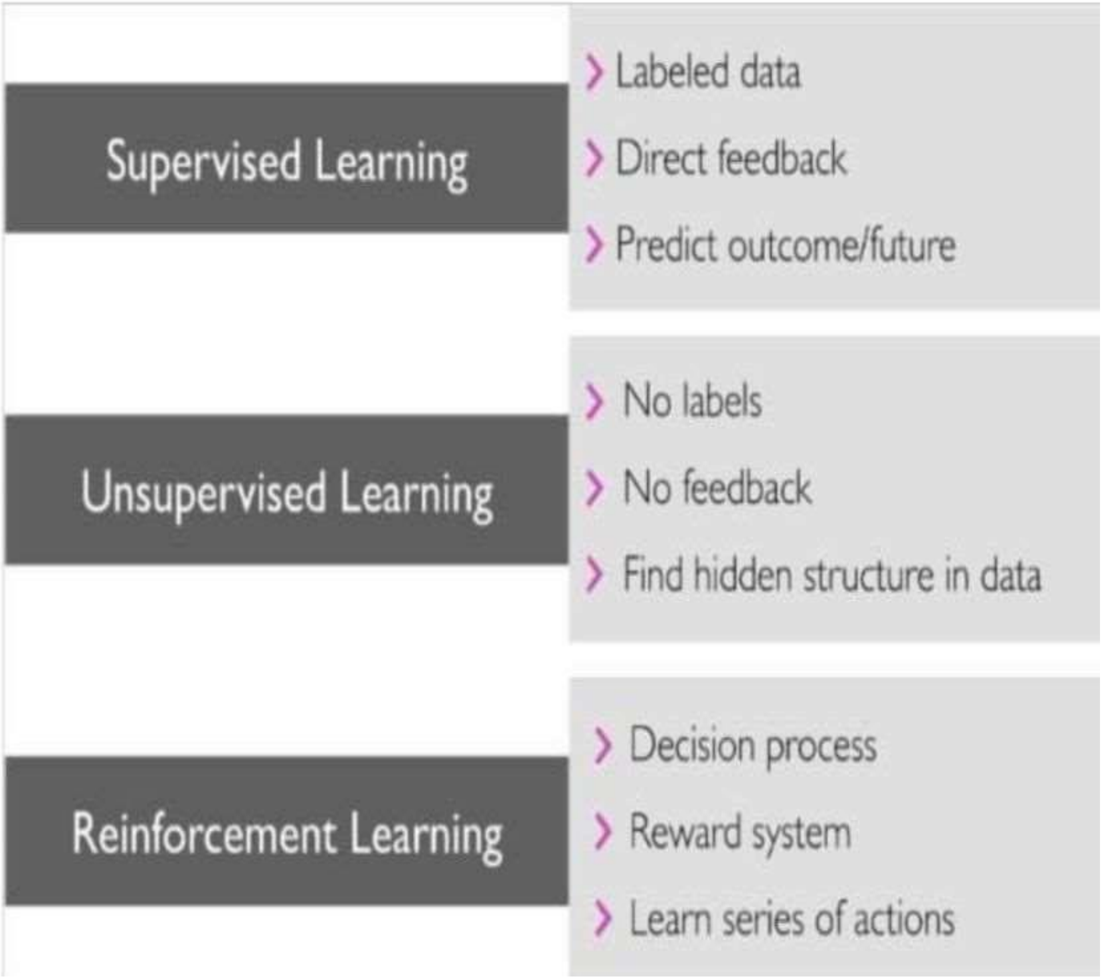
The dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

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

MACHINE LEARNING

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

However, training has to be done first with the data associated. By filtering and various machine earning models accuracy can be improved



Supervised Learning	› Labeled data › Direct feedback › Predict outcome/future
Unsupervised Learning	› No labels › No feedback › Find hidden structure in data
Reinforcement Learning	› Decision process › Reward system › Learn series of actions

Types of Machine Learning

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 a 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.

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.

Reinforcement Learning

Reinforcement learning is a class of machine learning which is concerned with how software agents ought to take action 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.

DESIGNING AND IMPLEMENTATION

Data Preparation & Cleaning

The data has been imported from the Kaggle website. The website provides a variety of data and the data used for the project is 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

for the project. The data was in a structured format and was stored in a CSV file format. The data was imported using the panda's 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.

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 modeling 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 the best-performing model (with thebest-performing parameter settings for each model) was selected.

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 the model by using different algorithms, different features, and different train test split sizes. 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

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.

The 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.

IMPACT ON SOCIETY

The following business ideas can be implemented in society which can benefit it as well:

- Increase trust and revenues by demonstrating positive social impacts.
- Improve worker productivity through responsible social management.

- Reach more customers who prefer goods and services that create social value.
- Promote dialogue with communities by preventing social conflict.
- Attract capital through profit-making with a social purpose
- Address social challenges to drive sustainable growth
- Stay ahead of new regulations on business and social impact.
- Spur innovation through social impact business models

CONCLUSION & FUTURE SCOPE

Background 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 premiumsfrom these models were compared with actual premiums to compare the accuracies of thesemodels. It has been found that Gradient BoostingRegression model which is built upon decision tree isthe best performing model.

Various factors were used and their effect on predicted amount was examined. It was observedthat a persons 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 a person's own health rather than other company's insurance terms and conditions. The models can be applied to the data collected in the 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.

LINK REFERENCES

https://www.moneycrashers.com/factors-health-insurance-premium- costs/

https://en.wikipedia.org/wiki/Healthcare_in_India

https://www.kaggle.com/mirichoi0218/insurance

https://economictimes.indiatimes.com/wealth/insure/what-you-need-to-know-before-buying-health-insurance/articleshow/47983447.cms?from=mdr

https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-tatistics.php

https://www.zdnet.com/article/the-true-costs-and-roi-of-implementing-

ai-in-the-enterprise/.

https://www.saedsayad.com/decision_tree_reg.html

http://www.statsoft.com/Textbook/Boosting-Trees-Regression- Classification