Health Insurance Cost Prediction

By,
Sameer
Yasin
Chachiya

Table of Contents

1) Introduction

1.1 Overview :

Health insurance is a type of insurance product that specifically guarantees the health costs or care of the insurance members if they fall ill or have an accident. Broadly speaking, there are two types of treatments offered by insurance companies, namely inpatient (in-patient treatment) and outpatient (out-patient treatment).

Generally, the one-factor-at-a-time method is used in experimental designs to determine the Charge properties. The major disadvantage of this approach is that it does not consider the interaction between the factors (interaction terms). The higher the number of the controlled and uncontrolled effect variables that influence the cost prediction,. Despite this, a few experimental designs have been suggested by considering the controllable effect variables and interaction terms between them.

1.2 Purpose:

In recent years, the ML methods have become popular as they allow researchers to improve the prediction accuracy of concrete properties and are used for various engineering applications. The ML methods have been used to increase the prediction accuracy of concrete properties, and the data derived from the literature sources were used

Regression models tend to be used for the prediction of the Charges or yhe Premium which is supposed to be paid . These models also demonstrate how they are related.

Previous studies evaluated the amount of the Premium to be paid by the customers and compared their results to the published data. In this study, the ML regression methods were compared to predict the Premium. The study aimed to determine the most successful regression method by comparing the random forest and Linear Regression.

2) Literature Survey

2.1 Existing Problem

This is generally determined when there was a time when the customers were not ready to pay for the premium. They thought the cost of the premium was high and was not worth it. Thus have a scarce of knowledge and resources the customers were less and thus this had to change and we came up with a solution.

2.2 Proposed Solution

The solution was made possible as we had a lot of data , having data can make a huge difference . using the dataset we can find out the co-relation among all the values in the dataset and thus use those approximations and find out the resultant prediction which determines what value . Thus, using this the customer will no longer have to be afraid of the premium charges as he can check his needs well in advance.

3) Theoretical Analysis

3.1 Hardware / Software designing

Python, Python Web Frameworks, Python for Data Analysis, Python For Data Visualization, Data Pre-processing Techniques, Machine Learning, Regression Algorithms

4) Experimental Analysis

The Premium charges data for the present work was obtained from the experiments. For generating a reliable data bank on Health cost, we had considered Six parameters, namely, age, sex, bmi, children, smoker and regionin the experimentalprogram.

Range of various parameters

Age – Range of Integer values.

Sex – Determing the Gender

BMI - For most adults, an ideal BMI is in the 18.5 to 24.9 range.
If your BMI is:

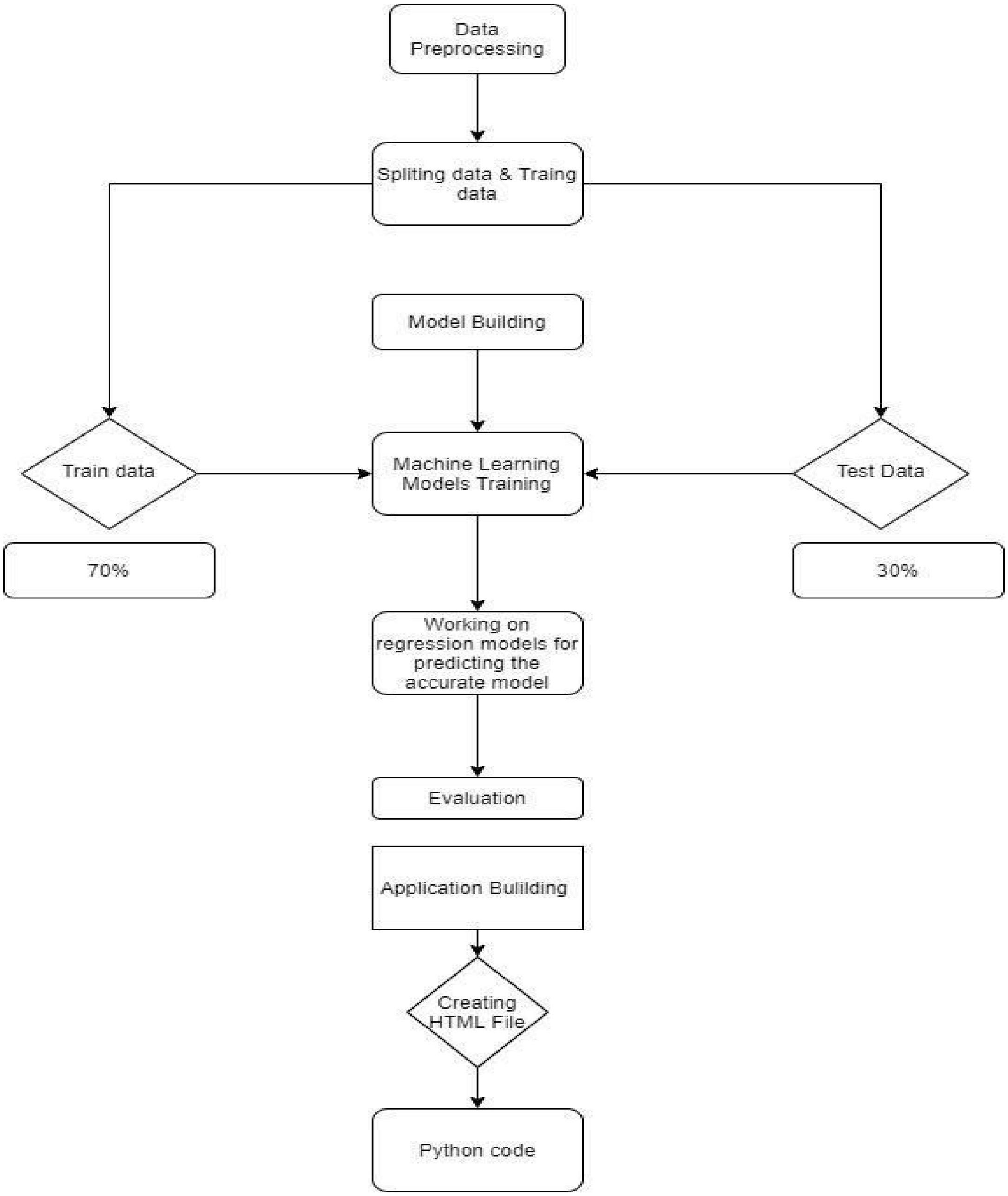
below 18.5 – you're in the underweight range
between 18.5 and 24.9 – you're in the healthy weight range
between 25 and 29.9 – you're in the overweight range
between 30 and 39.9 – you're in the obese range

Smoker - Smoker or Not

Region - region where they are situated

- Southwest
- Southeast
- Northwest
- Northeast

5) Flowchart


Data Preprocessing
Spliting data & Traing data
Model Building
Train data
Machine Learning Models Training
Test Data
70%
30%
Working on regression models for predicting the accurate model
Evaluation
Application Bulilding
Creating HTML File
Python code

6) Result

We have analysed the Insurance Data and used Machine Learning to Predict the Premium charges. We have used Linear Regression and its variations, Lasso, Ridge and Random Forests to make predictions and compared their performance. Random Forest Regressor has the highest accuracy and is a good choice for this problem. Random Forest Regressor trains randomly initialized trees with random subsets of data sampled from the training data, this will make our model more robust

7) Advantages and Disadvantages

Advantages:

Using Machine learning to predict the Premium charges will be time efficient and will give more accuracy in predicting the approximately close value can be done easily. It's more trustworthy and cost effective .It also helps in regaining trust amongst the firm and the customers

Disadvantages :

There is a 14% chance that the outcome will not predict the approximate value in that situation it can be troublesome.

8) Applications:

- Can predict the Premium charges using the inputs provided.
- Implementable on the website

9) Conclusion

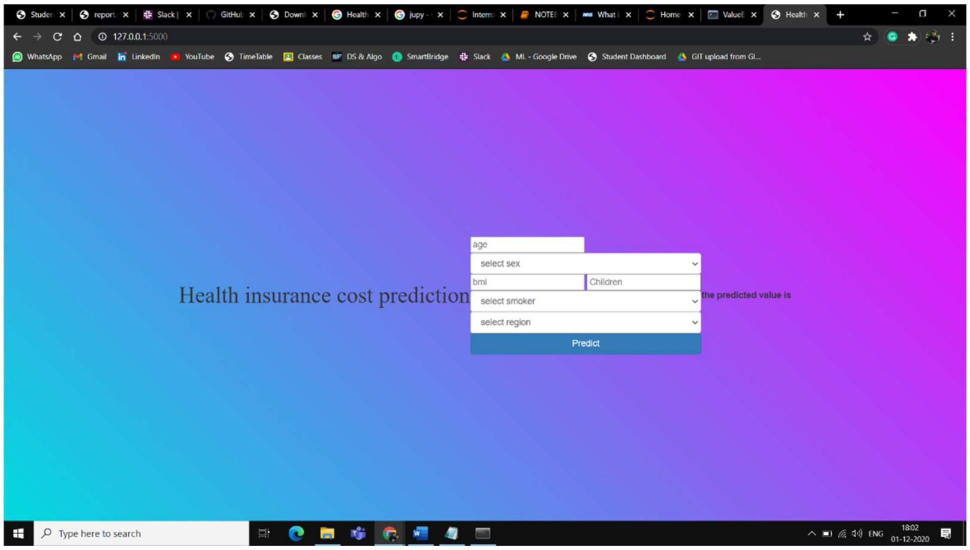
- Compared to all other Machine Learning Models Random Forest was best suitable for this data.
- Random Forest Regressor gave the maximum accuracy when tested using r2_score confusion matrix.
- Maximum accuracy received is 84.5%.

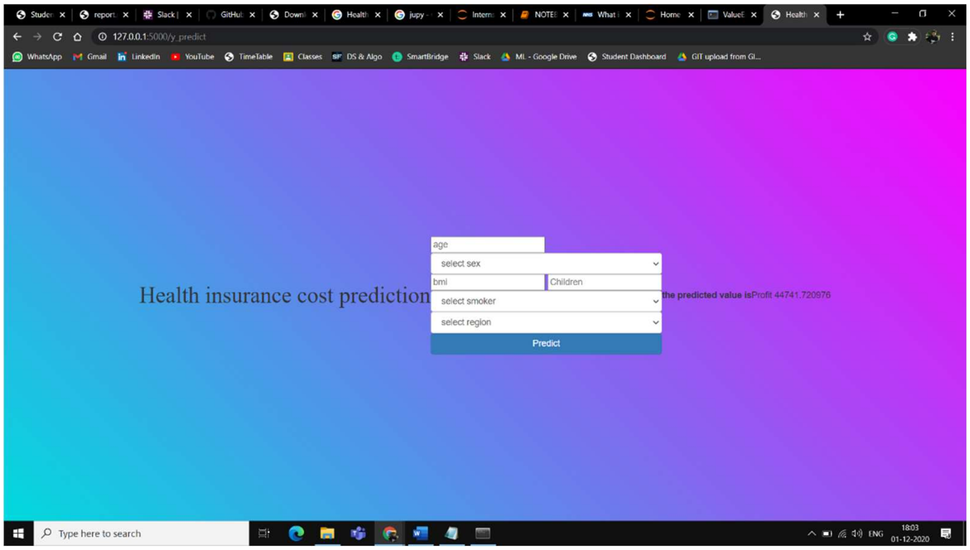
10) Future Scope

This model can predict the outcome with many different inputs within seconds. The model will save a lot of time for the Insurance Premium Companies. Experiment cost is also reduced which creates a bigger opportunity for these companies in cost effectiveness work.

11) Appendix

Screenshots





the predicted value isProfit 44741.720976