Medical Insurance Cost Prediction

Project report submitted to the Amrita Vishwa Vidyapeetham in partial fulfilment of the requirement for the Degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

Adithya Krishna - AM.EN.U4AIE21005

Adithya S Nair -AM.EN.U4AIE21006

Anoop Boby Manuel - AM.EN.U4AIE21015

Athul Gireesh - AM.EN.U4AIE21020

Navneeth Krishna - AM.EN.U4AIE21047

JULY 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AMRITA VISHWA VIDYAPEETHAM
(Estd. U/S 3 of the UGC Act 1956)
Amritapuri Campus
Kollam -690525

DECLARATION

We, Adithya Krishna(AM.EN.U4AIE21005), Adithya S Nair(AM.EN.U4AIE21006), Anoop Boby Manuel(AM.EN.U4AIE21015), Athul Gireesh (AM.EN.U4AIE21020), Navneeth Krishna (AM.EN.U4AIE21047) hereby declare that this project entitled Medical Insurance Cost Predictions is a record of the original work done by us under the guidance of Gopakumar. G, Georg Christoph Gutjahr Dept. of Computer Science and Engineering, Amrita Vishwa Vidyapeetham and that this work has not formed the basis for any degree/diploma/associations/fellowship or similar awards to any candidate in any university to the best of our knowledge.

Place: Amritapuri Date:
12-7-2022

Signature of the student

Signature of the Project Guide

Content

Introduction

A health insurance company can only make money if it collects more than it spends on the medical care of its beneficiaries. On the other hand, even though some conditions are more prevalent for certain segments of the population, medical costs are difficult to predict since most money comes from rare conditions of the patients. The objective of this article is to accurately predict insurance costs based on people's data, including age, Body Mass Index, smoking or not, etc. Additionally, we will also determine what the most important variable influencing insurance costs are. These estimates could be used to create actuarial tables that set the price of yearly premiums higher or lower according to the expected treatment costs. This is a regression problem.

Dataset

This dataset is taken from Kaggle. In this dataset, there are six independent variables and one dependent variable which are:

Independent: Age, Sex, BMI, Children, Smoker, Region

Dependent: Charges

Age: Age of the primary beneficiary (Min: 18, Max: 64, Mean: 39.22)

Sex: Insurance contractor gender, female, male

BMI: Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9 (Min: 15.96, Max:49.06, Mean:30.58)

Children: Number of children covered by health insurance, number of dependents smoker: smoking or not (Min: 0, Max: 5, Mean:1.09)

Region: The beneficiary's residential area in the US, northeast(Region 4), southeast(Region 2), southwest(Region 3), and northwest(Region 1).

Charges: Individual medical costs billed by health insurance (Min: 1121.87, Max: 49577.66, Mean:13030.00)

The shape of our dataset is 1338,7 where 1338 data are available for 7 attributes mentioned above

In the dataset given Sex, Smoker and Region are categorical values. In which Smoker is denoted by 1 and non-Smoker by 0, in Case of sex Female is denoted by 0 and male is denoted by 1.

Methodology

We use ten different models which predicts

- Age Vs Charge
- Sex Vs Charge
- BMI Vs Charge
- Children Vs Charge
- Smoker VS Charge
- Region VS Charge
- Age, Sex VS Charge
- Age, Sex, BMI VS Charge
- Smoker, Region, Children VS Charge
- Age, Sex, BMI, Children, Smoker, Region VS Charge

For evaluating these ten models we have used four criteria along with plots for two regression models which are Linear regression and Random forest regression

Four criteria are:

- **Mean Squared Error :**

The mean squared error (MSE) **tells you how close a regression line is to a set of points**. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs. To find the MSE, take the observed value, subtract the predicted value, and square that difference. Repeat that for all observations. Then, sum all of those squared values and divide by the number of observations. Notice that the numerator is the sum of the squared errors (SSE), which linear regression minimizes.

- **Mean absolute error** :

Mean absolute error (MAE) is a metric used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values. Here, errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable. They are calculated as follows :

$$\text{MAE} = \frac{|(y_i - y_p)|}{n}$$

- **Root Mean squared error :**

Root Mean Squared Error (RMSE) is a metric used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values. Here, errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable. They are calculated as follows :

$$\text{RMSE} = \sqrt{\frac{\sum (y_i - y_p)^2}{n}}$$

y_i = actual value
y_p = predicted value
n = number of observations/rows

- **R2 Score** :

R-Squared (R^2 or the coefficient of determination) is **a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable**. In other words, r-squared shows how well the data fit the regression model

The plottings were done between the Y test set and the Y prediction set for Linear regression and Random forest regression.

Fitting the Data into regression models

Importing libraries:

There are several libraries we are going to import and use while running a regression model up in python and fitting the regression line to the points. We will import pandas, numpy, metrics from sklearn, LinearRegression from linear model which is part of sklearn and r2_score from metrics which is again a part of sklearn

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import r2_score
```

Reading the Dataset :

We can use the read_csv() method to read the mpg dataset which we have into a CSV format at the working directory of python. The file is stored on a variable named URL, and this variable is assigned to the raw dataset taken from GitHub.

Cleaning of Dataset :

The data that we Imported needed to be cleaned to get precise Prediction. It uses functions inbuilt in Python. Some are :

- Data.info() - to get the information related to the data, the rows, Columns, The values present etc.
- Data.isnull().sum() - to get if there is any null values present in the data.
- Data.describe() - Used to describe the dataset (Count , Mean , min , max)
- There were 3 categorical values present in the dataset which are ['sex', 'smoker', 'region']. We converted them into numerical values using Label encoders.

Prediction of Medical insurance using various attributes

- We use 2 regression algorithms to predict the insurance prices which are Linear and Random Forest Regression. For the 10 models created we implemented the 2 algorithms and got the MSE, RMSE, MAE, and R2 scores.
- For which we used the train-test _split from sklearn to split the data into training and testing data.
- 2 variables are used, X and y as the 2 axes, and for the X value, according to which constraints are used, we will drop other independent variables from the list and for the y value we use the dependent variable which is “charges” to be predicted.
- The training data and the testing data are split into 80/20. 80% for training and 20% for testing and set the random state to 42.

- Fit the X_train and y_train into linear regression and Random Forest regression algorithm and use it to predict the X_test to get the desired y value using the predict method.
- Then we use the predicted value and the original y_test to get the MSE, MAE, RMSE and R2 scores.
- Plot the graph between the y_test and predicted values for linear regression and random forest regression.
- Prediction of the insurance cost using inputs given and prediction of the cost by using linear regression and Random forest regression

The performance of the Models can be evaluated by using the R2 score present, which intakes two parameters (y_test, predicted y values).

R2 score can be calculated using the formula => **1-(RSS/TSS)**

RSS = Residual Sum of Squares

TSS = Total Sum of Squares

If the R2 value is less than 0.4 we can say that the correlation between the parameters is less and if it is above 0.7, it is acceptable and above 0.9 is a high correlation.

Result

Through this experiment, we tried to predict the cost of the Insurance coverage to avail a patient. For the parameters, we took (Age, Sex, BMI, Children, Smoker, and Region)we got various R2 scores and could predict the cost based on the input.

- For Age vs Charges, the R2 score we got is:

```
The R2 score for linear regression 0.124089735395019 44
The R2 score is for Random Forest regression 0.0834335705680126
```

- For SEX vs Charges, the r2 score is:

```
The R2 score for linear regression 0.00261212606335659
The R2 score for Random forest regression 0.0027381382719894 32
```

- For BMI vs Charges , the r2 score is:

```
The R2 score for Linear regression is  0.039701931179418 78
The R2 score for Random forest regression is -0.3668672722790 2094
```

- For children vs Charges , the r2 score is:

```
The R2 score for Linear regression is  0.0016954628730256882
The R2 score for Random forest regression is 0.006294693468645 64
```

- For Smoker vs Charges, the r2 score is:

```
The R2 score for Linear regression is  0.6602486589056528
The R2 score for Random forest regression is 0.6598701889877197
```

- For Region vs Charges, the r2 score is:

```
The R2 score for Linear regression is  -0.0009124505146014261
The R2 score for Random forest regression is 0.010731190626061982
```

- For Age, Sex vs Charges, the r2 score is:

```
The R2 score for Linear regression is  0.12988839893083326
The R2 score for Random forest regression is 0.05531812368882416
```

- For Age, Sex, BMI vs Charges, the r2 score is:

```
The R2 score for Linear regression is  0.15590990487003253
The R2 score for Random forest regression is -0.0035097552981782076
```

- For Smoker, Region, Children vs Charges, the r2 score is:

```
The R2 score for Linear regression is  0.6635054447646268
The R2 score for Random forest regression is 0.6547483007887615
```

- For Age, Sex, BMI, Children, Smoker, Region vs Charges, the r2 score is:

```
The R2 score for Linear regression is  0.7833463107364539
The R2 score for Random forest regression is 0.863332812192262
```

Observations /Discussions

- Some of the single input data regressions have negative and lower r2 scores, it can be due to the noise in the data taken or can be due to overfitting.
- For AGE, SEX, BMI, and REGION the regression models show a very low r2 score.
- A low R-squared value indicates that **your independent variable is not explaining much in the variation of your dependent variable.**
- The single variable cost prediction is not acceptable as it causes a low r2 value and is sometimes negative.
- In some cases, the Logistic regression for a model works better than the random forest regression.
- As the count of input independent variables increases, The r2 score is also increasing and can have a better prediction or outcome.

Appendix

The colab Link for the whole code:- Colab