Analysis Health Related Expenses to Age, BMI, number of Children, and Smoker Status

Overview

Analysis fcused on health expenses on cohort of individuals. Dependencies were # of children, BMI, age and Smoker or not. Insights were generated individuals health expenses on these dependencies. Finally a LogisticRegression(LR) with solver at lbfgs, classification ML model was trained to predict an individuals propensity for health costs above $40,000 to asses for risk on the dependencies.

Findings:

1. Summary Statistics for Medical Charges data set
   1. Total rows: 1004 (=data points= individuals)
   2. Units assumed: Age: years; BMI in kg/m2; Charges: US$:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Units | Mean | Std | Min | Median | Max |
| Age | Years | 39.30 | 13.97 | 18.00 | 39.00 | 64.00 |
| BMI | kgm-2 | 30.79 | 6.19 | 15.96 | 30.50 | 53.13 |
| Children |  | 1.10 | 1.19 | 0 | 1 | 5 |
| Charges | US$ | 13,408 | 12,146 | 1,122 | 9,440 | 63,770 |

* 1. Categorical:

|  |  |
| --- | --- |
|  | Value-Counts (numbers of individuals) |
| Sex | Male: 508; Female: 496 |
| Smoker | Smoker: 312; Non-Smoker: 791 |
| Region | NE: 225; NW: 248; SE: 276; SW: 255 |

* 1. Graphical user interface

     Description automatically generatedScatter Matrix :
  2. Distributions:

|  |  |  |
| --- | --- | --- |
| Gender | Region | Smoker |
| Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated |

1. The sample that represents the population shows the following characteristics:
   1. Individuals in terms of numbers are distributed evenly, females to males’ demographics and between 4 regions [see above].
   2. ~1/5th of the individuals are smokers [see above].
   3. There are 3 distinct classes (bands) of individuals, as seen in the Insurance Charge vs. Age plot. The lowest band (green) is the cohort cheapest to serve, while the one on top (blue) is the costliest. [see below]
   4. There are 2 distinct classes (bands) of individuals, as seen in the Insurance Charge vs. BMI plot. The lowest flat band (orange) is the cohort cheapest to serve while the band gradually rising (red) behaves as generally expected (higher the BMI, higher the medical costs). [see below]

|  |  |
| --- | --- |
| Age vs. Insurance Charge | BMI vs. Insurance Charge |
|  |  |

* 1. The distribution of insurance charges shows skewed data set with few very high outliers (above 75% percentile). BMI has few outliers, while age has no outliers. They are both distributed evenly to the median. [Charges and BMI plots are seen below]

|  |  |
| --- | --- |
| Charges box and whiskers | BMI box and whiskers |
|  |  |

* 1. On average non-smoking males in the southwest has the lowest medical charges at $7,251, while the most expensive were smoking males in the southwest with charges at $35,230. [on the pdf code]
  2. The highest charge on an individual on average per region is for southwest with $14,953. [see below]
  3. On average the highest cost per individual depending on the number of children occurs at 2 children ($15,312) and the lowest ($8976) at 5 children. However, the latter case is suspect because of lack of data (numbers) for individuals with 4 and 5 children, as shown in the scatter matrix plot above. [see below]

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

1. The average age of the female is 40.00 years while for males it is 38.60 years. [see below]

Table

Description automatically generated

1. The smoking rate for those of who have kids = (47+45+26+2+1)/(186+47+151+45+94+26+14+2+11+1) = 20.97%

Smoking rate for those without kids = 92/(335+92) = 21.55%

Therefore, it is higher for those with kids. [see below]

Table

Description automatically generated

1. There is very low collinearity between variables (age, BMI, number of children) given by the correlation values and the heat map. [see below]

|  |  |
| --- | --- |
| A picture containing background pattern  Description automatically generated | Table  Description automatically generated |

1. a. Being male increases medical costs by on average $ 1,515. [see below]

|  |  |
| --- | --- |
|  |  |

b. Being a smoker increases medical costs significantly by, on average, $ 23,364. [see below]

|  |  |
| --- | --- |
|  |  |

c. Each additional year adds medical costs by $ 261. But this has a significant RMSE of $2,682.54. I’ve used average charges per year to reduce noise in this calculation as such there maybe some error associated, compared to raw data. [see below]

|  |  |
| --- | --- |
|  |  |

*Predicting Health Expense Risk from the Dependencies*

A predictive classification ML model was used where a patient’s healthcare cost above $40,000 being high risk (likely should be excluded). I’ve selected $40,000 as the cutoff because this is an imbalanced data set and will get highly imbalanced if the cutoff is set at $50,000. At $50,000, only 4 individuals would place themselves above the cutoff. At $40,000, the number of individuals above cutoff gets increased to 59. To mitigate this, we could use imbalanced learn classification ML algorithm such as one of the oversampling techniques: SMOTE or hybrid such as SMOTTEEN.

Data sets were separated: train (75%) and test (25%) using Scikit-Learn, train\_test\_split. They were then scaled using StandardScaler. LogisticRegression(LR) model with solver at lbfgs (Limited Memory Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS)) was used for modeling the system. Predicted values obtained and compared them with test (y\_test) values, and obtained LR classification report, the accuracy score is a respectable 98%. Also, the Precision and Recall scores were reasonable.

Of particularly important is the high-risk [1] recall score at 81%. This compared to Precision scores for both cases (high-risk[1] and low-risk[0]) is not as good. Since Recall scores have False Negatives (FN) (high-risk folks identified as low-risk), there is room for improvement. Of the test set (True 235 Low-Risk and True 16 High-Risk) of the low-risk predictions, 3 were misidentified as low risk (FN, which is costly to say an insurance provider) of the predicted high-risk 2 were True low-risk (False Positives FP), i.e., provider would drop them by which they’ll lose revenue (profit too since they are True low risk). But the actual financial losses are significant with the high-risk cohort, which means the model need to improve the Recall Scores. These metrics are seen on the Classification Report and Confusion Matrix.

Once trained, if an individual (or a cohort) applies, what must be done is to get the parameters for the individual (or the cohort) and see what the predicted risk (y\_pred from the code). As it stands, if the cutoff is at $40,000, the model would have an accuracy score of 98%.

A picture containing text, receipt, screenshot

Description automatically generated