Alec Knobloch and Jordyn Iannuzzelli

ECE:5995 Data Mining Fall22

Final Project Report

December 9, 2022

What Influences Higher Medical Insurance?

Abstract

The objective of this project was to identify ways to decrease medical insurance costs. Health insurance protects against unexpected medical emergency costs which, according to healthcare.gov, costs 9.12% of a worker's paycheck on average. Emergencies, pre-existing health conditions, lifestyle choices, and location could all have an effect on why the medical bills are high or low. Discovering what aspects increase medical insurance bills can help prevent or to plan for health insurance bills in the future. Our findings reveal attributes that could cost patients up to $20,000 more than average per year .

Introduction

The dataset selected consists of 7 main attributes: Age, Sex, BMI, Smoker, number of Children, Region, and Charges. For preprocessing aspects of the data cleaning, all data was placed into bins, which promoted more conclusive results. The ages ranged from 18-64 and were placed into five bins. “Sex” was divided between male and female. The smoker attribute was divided between yes and no. The number of children attribute ranged from 0-5 kids, and the region attribute was divided into northwest, northeast, southwest, and southeast of the United States. Data testing against the charges attribute determined which specific attributes affected the overall medical insurance’s charge total. To do in depth analytical analysis it was required to use numerical data, however attributes such as smoker region and sex were non-numeric. To fix this

issue, a second dataset was created which, for example, instead having male and female possibilities, sex was changed to 1 and 0. This was repeated for the other non-numeric data.

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
age	<28	28-37	38-47	48-57	57<	
sex	male	female				
bmi	<24	24-32	32.001-41	41.001-50	50<	
smoker	yes	no				
charges	<11622	11622-22122	22123-32623	32624-43124	43125-53625	53625<
region	southwest	southeast	northwest	northeast		
children	0	1	2	3	4	5

Figure 1: Binning for each attribute

Methodology

Research was conducted for this experiment using various data mining techniques such as averaging, discovering associations, feature importance, and modeling. First, data for each possible outcome were averaged to give a rough estimate of the possible correlations between which possibility had a higher influence over the total health insurance charge. These data were then graphed to illustrate and visualize any common trends. Taking into consideration possible outliers, these data were not enough to discover which attribute had a greater effect on the medical insurance charge. With the use of feature importance on a linear regression model, important attributes which correlated to the cost of health insurance were distinguishable. The regression model allowed a single attribute (dependent variable) to be scoped across another attribute (independent variable). The feature importance function took in a random number of decision trees (extra trees) and attempted to predict outcomes of various sub-samples. The results were averaged to derive the model's predictive accuracy. Alternatively, to determine definitive

correlations in the data, the research found well-supported association rules. This technique was used to cross-reference that the results of the regressive model were conclusive.

Experiments

Based on the methodology used to research what affects health insurance costs, the results were consistent across all the techniques. The regressive model showed that Region, Sex, and Children had little to no effect on the total charge due to the fact that either each option averaged at about the same cost, or there were no positive or negative correlations. Although some regions had a higher average than other regions, it was too close to draw a conclusion from this data. On the flip side, Age, Smoker, and BMI all had a positive correlation to the overall effect on total charge. From the averages of charges per age bin, it is determined that as age increases, total insurance charges increase. Average charges per BMI bin also recognized that as the BMI increases, the total average charge increases as well. As for smokers, those who smoke have a total average cost of $32,050.23, and those who do not smoke have a total average of roughly $8,434.27.

Figure 2: Relative averages over non-directly correlated attributes

Figure 3: Relative averages over directly correlated attributes

The feature importance on regression with extra trees recognized that the order of least importance was Sex, Region, Children, Age, BMI, and Smoker.

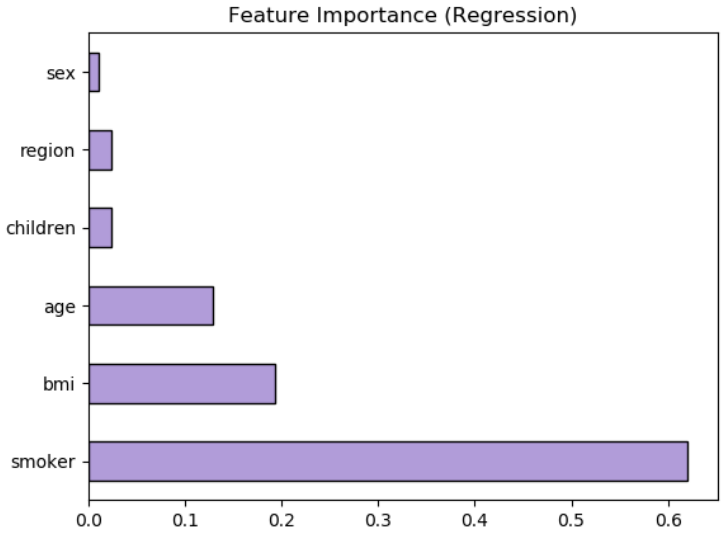

Figure 4 Feature importance for each attribute

Results from modeling the dataset against sub-test sets, predicted charges were similar to the actual charges with a 0.79 accuracy. Mean absolute error equated to roughly $3954.18, mean squared error was roughly $31,987,105.24, and root mean square error was $5655.71.

Mean Absolute Error: 3954.1892639155058
Mean Squared Error: 31987105.236794207
Root Mean Squared Error: 5655.7143878376855

Accuracy: 0.7989875396812893

Association rules were used to compare if the data was consistent with common trends in the dataset. With a minimum confidence of 50% and a minimum support of 30%, the rules that stuck out significantly were charges $32,624 - $43,124 correlating to smokers, with 96% confidence. Adding onto said bin, charges of $43,125 - $53,625 led to a smoker of a confidence of 100%. This stayed consistent with the initial results showing that the Smoker attribute had the highest influence on the total charge.

SMOKER
Rule: charges=32624-43124 -> smoker=yes
Support: 0.07249626307922272
Confidence: 0.9603960396039604
==========
Rule: charges=43125-53625 -> smoker=yes
Support: 0.035127055306427506
Confidence: 1.0

BMI
Rule: bmi=33-41 -> charges=32624-43124
Support: 0.03886397608370702
Confidence: 0.5148514851485148

Conclusions

To conclude, the results from each test were consistent. The most important technique used to find the data results was feature importance. This technique gave a distinct weight for each attribute which helped display how each attribute could impact the total charge. As for the relative averaging technique, the data could have been skewed based on outliers and outside implications. The use of mean opposed to median in this case allowed for an overall scope of the data opposed to the center of the data. When conducting results based on the methods used, it became clear which attributes were most effective and least effective. Age, BMI, and Smoker all had the greatest effect on the overall insurance cost. Some attributes such as Children fluctuated between an important attribute and non-important attribute. The Children result was considered inconclusive due to the fact that the number of children had little effect on the overall feature importance, but the regression model showed a negative correlation.

Some of the strange discoveries in the results could have been affected by outside factors. For example, people above the age of 65 will receive medicare, which affects the overall insurance charge. The linear regression chart results for the Children attribute showed that as the number of kids increased, the health insurance costs decreased. A conclusion from this result could be that parents who have more kids may be less likely to smoke. Outside factors other than the given attributes could have an effect on other attributes as well such as medicare for the elderly, and subsidized insurance for those struggling with poverty.

Based on the results of this experiment, recommended future experimentation might include , a function using the derived weights from the feature importance function which would implement and test the data to create a make-shift model for this dataset. The use of this function would be to find an exact weight for each attribute option, aka male importance weight vs female importance weight. Although deriving the relative averages for each individual attribute option

showed clear correlations, outliers may have skewed the results and would not have given an effective weight to distinguish between the other options.