**Annual Individual Medical Charges Analysis**

Prepared by Carter Harris

3/31/2020

**Abstract:**

To understand a 1400 row dataset on individual annual medical charges, a series of predictive models were studied. These models were used to study the effects of each of the 6 predictors present in the dataset and reduce them to only the most effective and accurate for the task at hand. Then, these key predictors could be combined into the optimal model for predicting charges based on a given individual’s characteristics.

**Key Takeaways**

* Smoking causes far and away the highest rise in medical charges, with a smoker totaling roughly $24,000 more on average than a comparable non-smoker.
* Obesity was another major factor, with each additional point on an individual’s BMI raising costs by $365.
* Both are highly accurate predictors on their own but combining them into a single predictor resulted in a remarkably accurate model based upon common statistical metrics.

Due to various factors most of the modeling techniques yielded similar results, however the combined model mentioned above was unquestionably the most accurate option. Utilizing this model and its most important predictors we can then analyze potential new customers to find those who are likely to have the highest medical charges.

**Analysis Methodology**

To analyze charge trends within the dataset, I tested 6 different statistical modeling techniques to find the most accurate predictions. A key aspect of the best model is that it must achieve this while also not becoming excessively complex to the point where it could harm the model’s performance.

The dataset studied contained 7 different columns: age, sex, BMI, children, smoker, region, and charges. Out of the roughly 1400 rows, 75% is made into the training set for building the model with, while 25% is saved as a test set to evaluate the model. Charges was then used for the prediction target, while the others were set as prediction variables and then analyzed to see what effects they have on charges. Once these effects are established, the most significant and accurate of these can then be run through the models below to find the optimal version. Accuracy is checked through the Mean Absolute Percentage Error (MAPE), which works by finding the error between predicted values and the known validation data and then turning it into a percentage. For an example, 47.5 means that the model on average is off by 47.5%.

|  |  |  |
| --- | --- | --- |
| **Model Description** | **Important Predictors** | **Accuracy in MAPE** |
| Multiple Linear Regression | Age, BMI, smoker, children | 47.50 |
| MLR + dummy obesity variable | Age, BMI, smoker, children, obesity | 47.67 |
| MLR and obesity + smoker | Age, BMI, smoker, children, obesity+smoker | 25.95 |
| Exhaustive Search | Age, BMI, smoker, children, obesity | 51.41 |
| Forwards Selection Algorithm | Age, BMI, smoker, children | 47.50 |
| Backwards Selection Algorithm | Age, BMI, smoker, children | 47.90 |

**Commentary and Suggestions**

As mentioned previously, choosing the best model means considering both accuracy in predicting the test set and keeping complexity to a minimum to preserve model performance. The above models would all be useful to one degree or another, however, it is clear based upon the MAPE scores on the right that the MLR and obesity + smoker model is far and away the most accurate.

Using this model to predict the charges of new customers would undoubtedly bear dividends in being able to understand the potential risks they involve. This would help us to be able to adjust rates to the customers’ risk levels, as well as minimize unforeseen and costly risks down the road. Additionally, the knowledge gained here about the massive effects of factors such as obesity and smoking status should spur the company to look towards measures encouraging healthy behavior among customers as a method of simultaneously improving lives and cutting charges.

As it stands we have a relatively narrow range of data about each person, so expanding the available data moving forward would undoubtedly lead to even more accurate and therefore valuable predictions.