Brenden Hale

MIS 64060

**Data**

The dataset I am using contains 1,338 approved and released patient data records provided to Packt, a publishing company, from a partner. According to Packt, the location information was obscured for patient anonymity and instead uses the corresponding US Census region.

**Problem**

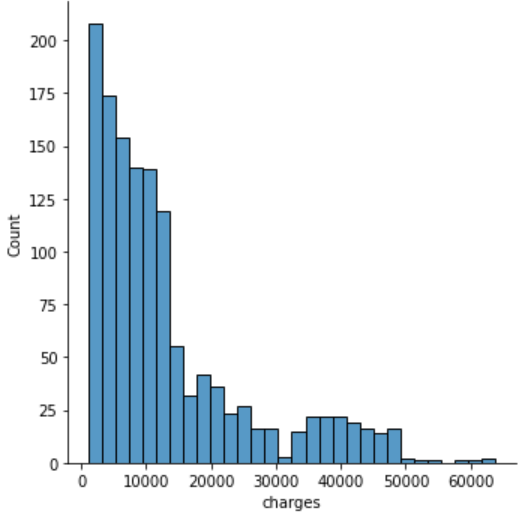
As a newcomer to the insurance industry, I quickly realized how dynamic it is. Although the overarching business model remains the same, there are countless variables behind the scenes that perpetually move. Typically, any medical expenditure is costly for both the beneficiary and the insurer. Given the nature of the business, insurance companies need to payout less than what they take in on insurance premiums. This is where the importance of prediction comes into play. Given a person's age, lifestyle, and other factors, insurance companies need to be able to predict what a specific individual may accumulate in medical expenses during the calendar year. Based on this information, they can fluctuate insurance rates as necessary.

**Approach**

To accurately predict a person's projected medical costs, we must create a machine-learning model. The first step is to gather real-world data that would allow us to develop a model based on a person's attributes, such as age and the charges accumulated over a calendar year. Given the requirements, this is a clear use case for linear regression. After evaluating and cleaning the data, I allocated 60% of the data to the training set and 40% to test the model. The next step is training the model with the data. I then verified the model's fit, which displayed a moderately high-value, portraying the model's competence in predicting medical expenses for a given person.

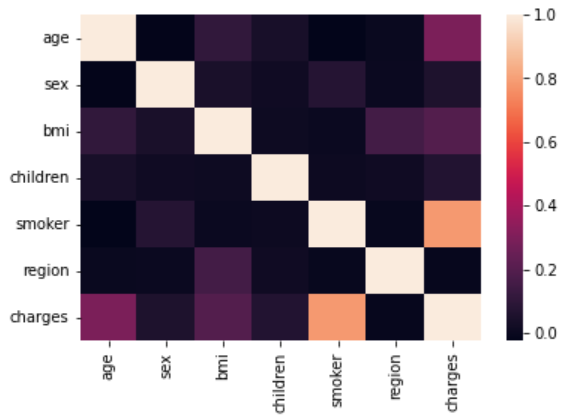
**Analysis**

During the data exploration phase of the project, it is evident that the charges in the dataset are skewed to the right. The vast majority of the accumulated medical expenses were between $1,100-$10,000. This is no surprise, as it is unlikely that every person would have undergone a major surgery.

****

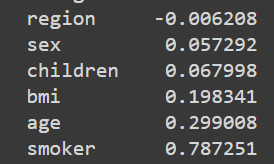
**Figure 1.** Histogram of charges vs. count of beneficiaries

Next, it is essential to categorize what characteristics in the dataset correlate with the medical expense impact. We created a heatmap to visualize this. We can then extract the data and rank which variables observed in each individual attract the most significant medical expense.



**Figure 2.** Heatmap of correlation between beneficiaries' attributes

This visualization depicts a substantial correlation between smoking and charges. Based on the data in figure three, being a smoker significantly affects medical expenses. Age and BMI also generate greater costs; however, it does not compare to those who smoke. Children, sex, and region play minimal roles in determining medical expenses.



**Figure 3.** Ranking attributes in relation to affect on charges

**Conclusions**

Using the model, we can predict the total medical expenses a customer may incur over a calendar year with a fair degree of certainty. This will allow us to create a dynamic and competitive pricing model while keeping our margins comfortable. The prediction technology will also enable us to serve our customers better by returning them quote information faster. The model performance can be adjusted based on the weight placed on certain characteristics such as smoking, age, or obesity. It can also be fined tuned with an enhanced dataset. This approach provides the business with flexibility in its prediction models.