Daniel Flores

2/20/2024

Final Report

Health Expense Prediction

# Introduction

One of the greatest examples of using data science is in healthcare. These companies use data science to try to predict trends and predict how much expenses someone might need. To do this, they use past data to try and predict those expenses. Someone that engages in risky behavior such as drinking or smoking would have a higher premium than those that didn’t. Your occupation also played a role in your insurance, as working on skyscrapers has a higher occupational hazard than working in an office. After gathering all this data on an individual, they use regression to predict the expenses of an individual, and then adjust the price of their healthcare accordingly. Of course, I can only predict those expenses. How the companies set the prices is beyond my realm of knowledge.

To mimic this expense prediction, I decided to see if I could use multiple regression using a variety of features to predict our target variable. After searching around on Kaggle, I found a dataset that fit what I was looking for. This dataset contains 1,338 entries and 7 features, one of those features being our target variable, the medical expenses of the individual. The rest were features that were, as I described above, different data that might affect how much an individual spent on health expenses. These features included the following about the individual; age, bmi, children, expenses, sex, smoker, and region. After seeing this dataset, I wanted to also create some visualizations for the dataset to get a better understanding of the data.

What follows is my report on the model built and the visualization I made that could help better understand the content of the data. It includes the processes I used to make sure my data was valid, as well as making sure that the data I used to train the model was ready to do so. I needed to be sure that the model was trained using the data correctly, or the model would fail. While creating the visualizations isn’t the main goal, I think it helps put the data into context and is good practice.

# Data

The original dataset consisted of 1338 rows and 7 columns. Of those columns, 4 were numerical, being age, bmi, children, and expenses. There were also 3 categorial columns, being sex, smoker, and region. I changed these to numeric values for the purpose of creating the model. For this model, I decided to omit region, as I didn’t want location to affect my model. The ‘age’ column refers to the patient’s age, as an integer. ‘bmi’ represents Body Mass Index, which indicates the amount of bodyfat on a person’s body, calculated by their weight in kilograms divided by the square of height in meters. ‘children’ represents the number of children the patient has. ‘expenses’ is their medical expenses. This is what I plan to use my model to predict. As for the categorial attributes, ‘sex’ represents the patient’s sex, 0 for female, and 1 for male. ‘smoker’ represents whether the patient smoked cigarettes, 0 for no, 1 for yes. So, using age, bmi, children, sex, and smoker, I wanted to predict expenses. The first thing I did was remove the region column. I also checked for null values in my dataset using .isnull().sum(). None of the rows or columns had null values. I also double checked the data types using .dtypes. Age is an integer, sex(categorical) is an integer, bmi is a float, children is an integer, smoker(categorical) is an integer, and expenses is a float.

To better understand the data, and to see any immediate correlations, I created a simple correlation heatmap. Doing this showed that smoking had the strongest correlation to Expenses. This will be important later when we want to understand correlation and start building the model. Next, I wanted to search for outliers. To do this, I calculated the mean and standard deviation of Expenses, since that is our target variable. The Mean was 13270.422414050823. The standard deviation was 12110.011239706457. With these numbers, I can calculate the upper and lower limits of the data. The following code was used to calculate those limits. The second output is the amount of data from my dataset that is in those limits, in percentage.

I then used those limits to create a separate data frame that only contained the entries within that limit. After doing this, I checked the shape of the data frame, and I found that I had dropped 7 entries using this method.

Going back to the region attribute that was dropped. The reason that I believe it was a good idea to remove region was because of its content. It was a categorical attribute, that had 4 unique regions, southwest, northwest, southeast, and northeast. The reason I decided to remove this feature from the model was because of its dispersion. All of the values were between 23 and 27 percent, showing almost equal representation. Because of this, I wanted to create a model that would predict regardless of region. I understand that this model may not be as accurate as possible, but I believe it could be used when predicting across the map for example.

The visualizations made for the dataset were each made with the purpose of showing the variety of each feature. For example, box plots were used to show the dispersion of data, while bar graphs were used in the case of children, to show how many entries had 1 child, 2 children, 3 children, and so on.

# Methods

All of my data cleaning and analysis was done through Jupyter Notebook. Imports as followed:

import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error

from scipy import stats

import statsmodels.api as sm

import statsmodels.formula.api as smf

from sklearn.preprocessing import LabelEncoder

I used the combination of these packages to complete this task. sklearn’s linear model is simple to use and can be used for multiple regression. Using pandas, I created a data frame using my dataset. From there, I was able to take a closer look at the data. I was able to bring up information such as mean, standard deviation, as well as the percentiles. I then used LabelEncoder to change some of the categorical data into numeric, for our model to be made. For sex, I made it binary, 0 for female, and 1 for male. For smoker, 0 means they are a non-smoker, and 1 means they smoke. As I mentioned above, I decided to drop the region feature from the data frame. After checking for null values, I was pleased to see that there were none. I then wanted to double check that my features were all numeric, and therefore ready for use in model training. After making sure of that, I knew I needed to address outliers. To do this, I decided to calculate upper and lower limits.

After doing that, I created a new data frame that consisted of only entries where the expenses were within the upper and lower limits. This removed any outliers. Once I knew I had the data that I was going to use in the model, I visualized the different features. This was done using a combination of seaborn and matplotlib.pyplot. With these packages I was able to make box plots, heat maps, and bar graphs to represent the data.

To create the model, I separated my features and my target variable. In this case, smoker, age, sex, bmi, and children were the independent variables, or the features. Expenses was my target variable, or my dependent variable. I then split my data to use some of the data to train my model, and then the other data to test my model’s accuracy. I decided on making my test size .2, as that is a agreed upon standard for model testing. After creating and training my model, I calculated the intercept, as well as slope of each feature.

# Analysis

I started the analysis with 1338 rows and 7 columns. After dropping the region column I’m left with 6 columns. I then converted the categorial data into numerical, for use in regression later. After doing this I was presented with the following output when I wanted to observe some statistics. A screenshot of a computer

Description automatically generated

When working with any data, it’s important to remember that outliers are always possible. Also there will always be factors that could affect the actual outcome that we don’t know about. For example, genetics aren’t considered in this test. That being said, we can see that while the average expenses is around thirteen-thousand dollars, the maximum is sixty-three thousand dollars. Within the context of this data, it is more likely than not that an independent variable outside of this data is responsible for the outlier.

Next, I wanted to be sure that my data was preprocessed and ready to be used to train my model. To do this, I double checked that there was no null values in my data. There were none. I also wanted to be sure that all of my data was numerical, and to do this I used .dtypes. They all were. Next, I wanted to get rid of some of the outliers I mentioned above. To do this, I decided to calculate the upper and lower limits, and only use the data entries that fall within that limit for expenses. The upper and lower limits are represented by being within 3 standard deviations of the mean. The mean of the expenses was 13270.422414050823, and the standard deviation was 12110.011239706457. We saw this in the figure above, but we needed them as variables to calculate the limits. A screenshot of a computer code

Description automatically generated

Using this code, I was able to find the upper limit of around $49,600.45. Lower limit won’t matter when removing outliers, since we can’t have negative payments on healthcare. We can also see that 99.5& of our data fell within our limits. The empirical rule states that 99.7% of data should fall within the calculated lower and upper limits. This data follows that rule pretty closely. After that, I created a new, final data frame that only included values from within those limits. After checking with .shape, I can see that we lost 7 rows of data, all of those being above that upper limit of $49,600.45. From here, I knew I had the final dataset to use for my model, but first, I wanted to crate some simple visuals, just to understand the demographic of the patients. A screen shot of a graph

Description automatically generated

To view some correlation before creating a model, I went ahead and created a heatmap that showed correlation.

A screenshot of a computer screen

Description automatically generated

It was obvious that smoking had the most affect on expenses from here. Next, I built the model. I used smoker, age, sex, bmi, and children as my features, and made expenses my target variable, since that’s what I want to predict. Here is the code I used to build my multiple linear regression, as well as ANOVA analysis of the model.

A screenshot of a computer

Description automatically generated

We can see that the r-squared value is .75. When it comes to r-squared values, the closer to 1 it is, then the more accurate the model is. Anything above .7 is considered high correlation. Because of this, I can say that our model is pretty good at predicting expenses. In order to understand each variable’s values, for example, we can see that smoker’s t-score is 57.75, and a p value smaller than 0.000. In general, the further from 0 the t-score is, the more the variable affects or target variable predictions. For p-values, the smaller and closer to 0, the more significant the variable is. In general, variables with t-scores further from 0, and p-values closer to 0, have the most statistical significance. To see how the predictions went, we can make a separate data frame that has our test data next to our predictions.

A screenshot of a computer

Description automatically generated

We can see that while some aren’t quite as close, most of these predictions are pretty close to our actual values for each entry.

# Conclusion

As I defined at the beginning of this report, my goal was to create a model that accurately predicts medical expenses, based on various dependent variables. The purpose of this is to mimic the practice used by healthcare companies. After cleaning and preprocessing the data, I was able to create a multiple regression model that predicts medical expenses of a patient based on the features of whether they smoked cigarettes, their age, their gender, their BMI, and how many children they had. The model had a r-squared value of .75, that 75% of the variance in the target variable can be explained by the independent variables. We can even use ANOVA to understand each variable’s statistical significance. In our model, whether they smoked had the most statistical significance to medical expenses. The next significant was age, followed by BMI, the number of children they had, and lastly, their gender. Realistically, this makes sense. It is well known that smoking is bad for you, so that makes sense to be the highest. What was interesting was how much more significant it was. This makes me wonder how accurate that is in the population. Next, age and BMI make sense. We know that being in better shape is best for your overall health and can prevent many hospital and doctor visits. Age also makes sense, as our health begins to slowly fade. Sex and the number of children were the least significant. Even after removing an attribute, we were able to create a model that predicted expenses with a significant r-squared value. It’s important to remember that there was this omission.

# Appendix

Dataset- <https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction>