Medical Insurance Cost Prediction: Machine Learning Project

October 08, 2021

A. Introduction

A1. Background

Rising health care costs are just one of the important problems everyone in the world has. The value of health insurance claims data in medical research has often been questioned because databases are designed for financial reasons and not for clinical purposes. The predictive power of claims data became a topic of research in the 1980s and numerous studies have since established the predictive power of administrative data on health-care costs.

Predicting the costs with great accuracy is an objective that a lot would want to solve, be it people in insurance or people who have medical needs. This highlights the need for tools that can accurately predict the costs of health insurance based on factors that are easily available. In this project, we would take a data-driven approach for the prediction of medical insurance cost using machine learning model. The aim of this model is to accurately predict the medical insurance cost of a patient depending on five attributes.

A2. Problem

For this project, the question that can be asked: “What Machine Learning Model can be used that can accurately predict medical insurance cost based on five factors?”

A3. Target

The intended audience for this project are people who work for insurance companies so they can have a basis on how much to charge each client. It can also be used by clients to have a general knowledge on how much medical insurance would cost them.

B. Data Acquisition and Cleaning

B1. Data Source

For this project, the following data is used:

1. Dataset that contains information about each client:

Data Source = Kaggle: https://www.kaggle.com/mirichoi0218/insurance

Description: The dataset contains 1338 entries and 7 attributes.

Matrix column entries (attributes):

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

B2. Data Distribution

Before doing the analysis, we thought it would be best to visualize the distribution of all attributes to get some insights about each feature.

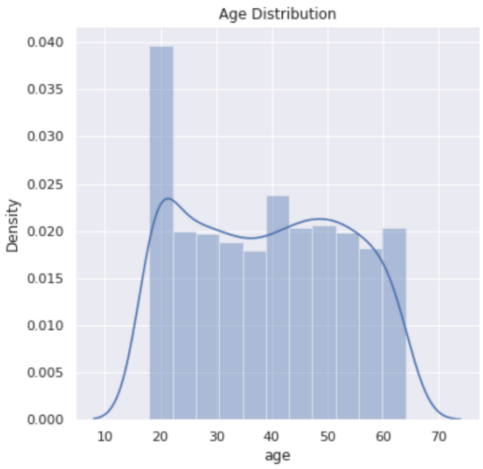

Figure 1: Age Distribution

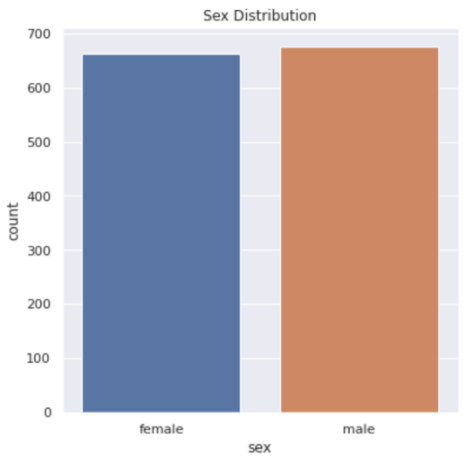

Figure 2: Sex Distribution

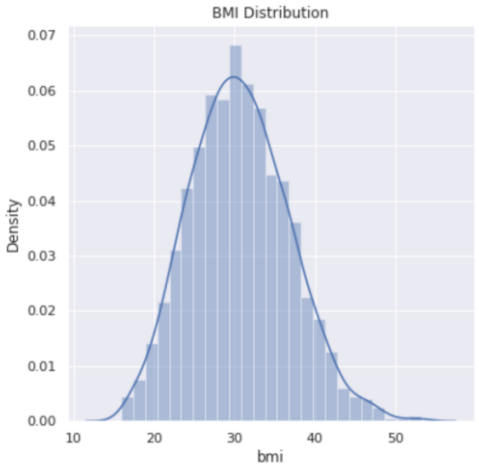

Figure 3: BMI Distribution

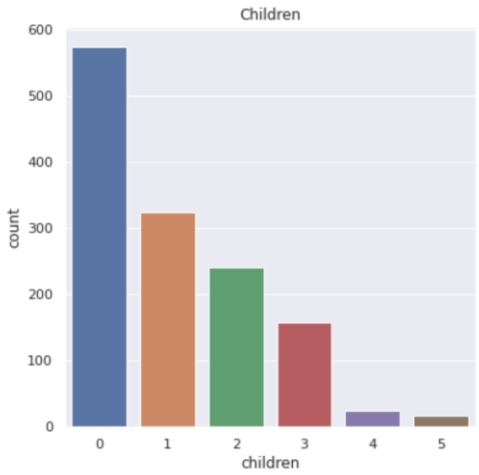

Figure 4: Children Distribution

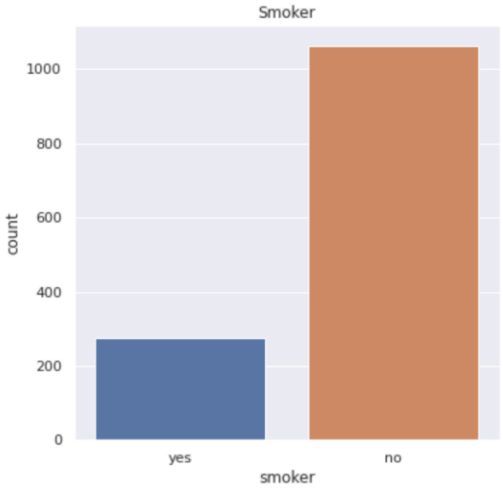

Figure 5: Smoker Distribution

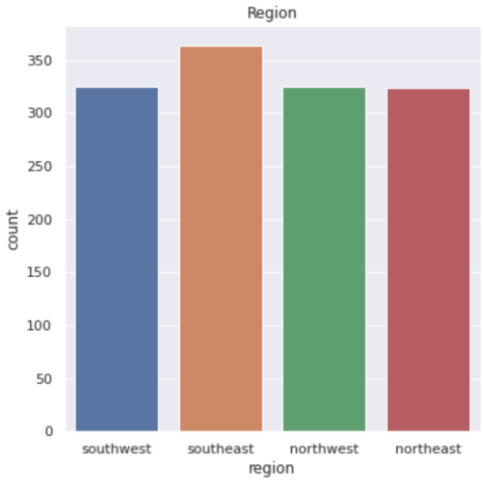

Figure 6: Region Distribution

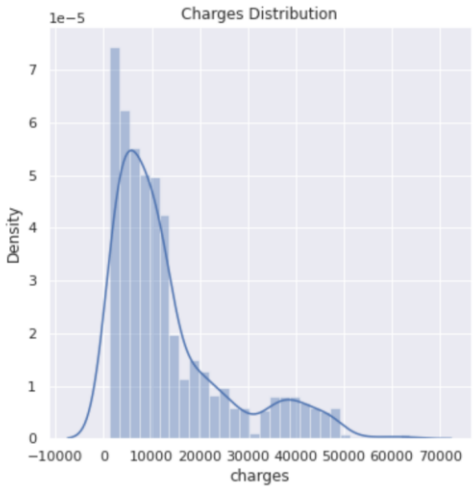

Figure 7: Charges Distribution

Looking at Figure 1, the age ranges from 10 to 70. The maximum density, or the most number of values, in the 19-22 range. The values around 23 or 24 to 70 have similar values or almost equal. Only in the age range 19-22 are there the highest number of beneficiaries.

Figure 2 shows the sex distribution, which shows a greater number of male beneficiaries. Specifically, there are 678 male beneficiaries while there are 662 female beneficiaries.

Figure 3 is the BMI distribution, where the bmi suggests whether a person is overweight or underweight. Looking at the distribution plot, it shows a normal distribution where most of the beneficiaries along the 30 range. We have similar distribution on either side of the peak. Since the normal BMI range is around 18.5 to 24.9, we can see that the highest densities are along 25 to 35 range there are many beneficiaries that are considered overweight. This might increase the insurance cost.

Figure 4 shows that most beneficiaries have no children, specifically 574 of them. While 324 have one child; 240 have two children; 157 have three children; 25 have four children; and 18 have five children.

Figure 5 shows that most beneficiaries are non-smokers while only a small portion are smokers. Using the value_counts() function, there are 1064 non-smokers and 274 are smokers.

Figure 6 shows that the highest number of beneficiaries come from the southeast region. Other regions have similar number of beneficiaries. Using the value_counts() function, there 364 that came form southeast, 325 from southwest, 325 from northwest, and 324 from northeast.

Lastly, Figure 7 shows the distribution of the charges attribute. This shows that there are a lot of data distributed around the 10000 values and little values around 30000 and 40000 range.

B3. Data Cleaning

The Kaggle Dataset initially have to be downloaded in the data assets column in Watson Studio. The file has 1338 entries and 7 attributes.

We initially took the information about the dataset. This is to see what type each attribute is and to see if there are attributes that have missing values. When ensured that there is none, we started the pre-processing stage.

As Machine Learning and Deep Learning models will be used for this project, we have to encode the values for the categorical features into numeric as the models require input and output variables to be of that type. We encoded the categorical attributes – sex, smoker, and region – into numeric values. For the ‘sex’ attribute, we gave male = 0 and female = 1; for the ‘smoker’, yes = 0 and no = 1; and for the ‘region’, southeast = 0, southwest = 1, northeast = 2, and northwest = 3.

Then the attributes were split into features and target, where all attributes except ‘charges’ were included in the features data and labelled X and the target data consisting of only

‘charges’ and was labelled Y. After, the X and Y data was split into training data and Test data where 20% of each data go to the Test data. The training data have 1070 entries and the Test data have 286 entries. This is the dataset to be used for data analysis.

C. Exploratory Data Analysis

For this project, the aim is to use a Machine Learning Model to be able to predict accurately medical insurance cost of each beneficiary. Hence, we will specifically use a Multiple Linear Regression Model as we want to calculate Y, or ‘charges’, from a linear combination of the input variables, or X.

To use this model, we imported the LinearRegression library from sklearn. This would fit a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. After the Linear Regression model in the notebook, we then fit a linear model using the X_train and Y_train datasets.

After fitting, we used the predict function on the X_train dataset to calculate charges based on the values in the dataset. To assess the accuracy of the prediction, the R-squared is calculated. R-squared is a statistical measure of how close the data are to the fitted regression line. Taking the R-squared value of the training prediction and comparing it to the actual training ‘charges’ values, we get a value of 0.75. As the value is high, this model suggests that the model is good. Another important metric is to predict ‘charges’ using Test data. Using the same process as for the training data, we get a R-squared value of 0.74. As the two R-squared values are quite similar, there is no overfitting issue present.

C1. Building a Predictive System

Now, we can use this Linear Regression model to calculate what the insurance cost should be for each beneficiary based on several factors. To do this, we took a row from the original dataset and deleted the “charges” value. We also encoded the categorical attributes to the proper values. This would be the input data. We initially changed the input data to a numpy array and then reshaped it to a one shape dimension. We did this as the Linear Regression model expects several input data. Then we used the model to predict the insurance cost based on the input data. While the output does not give the same exact charge as in the original dataset, it does give a value that is very close to it.

D. Conclusion

Medical insurance cost has been an important problem that many are facing. Being able to accurately predict how much the cost would be based on a person’s age, sex, region, and other factors is just one of the important issues an insurance company face. A linear regression model helps in this dilemma by using these factors to be able to predict the insurance cost. From this model, we can create a predictive system that can be easily available everywhere and easy to use by insurance companies as a guide to properly request from a beneficiary for their services.