Health Care Cost Prediction

Midterm Project Report for Data 1030

Yifei Song

October 11, 2021

1 Introduction

We spend part of our money on health care every year, so I think it is necessary to collect some data to predict the annual expenses according to each person's own situation, and then purchase appropriate medical insurance or save enough money to deal with the possible expenses according to the conclusion.

This project attempts to create a regression model from the health condition of an individual to predict his or her annual health care cost. The dataset used for this project came from the Medical Cost Personal Datasets in Kaggle, which includes age, sex, bmi, children, smoker, region and charges. Through 6 features of 1338 individuals, I want to explore which factors are closely related to our target variable 'charges', and whether we can predict some person's medical insurance expenditure once we obtain these factors. This will greatly help the insurance company price medical insurance. Furthermore, individuals will have the benefit of understanding the medical expenses he will pay as well.

Several authors have uploaded their study to kaggle that use this dataset to establish various models to predict the health care cost. Shubham Shrimant used Decision Tree Regression to do the prediction based on all 6 features. The model he made performed well for accuracy, with a r2 score of 0.69.[Shr]. Praviin Jaiswal used the sequential model in Keras multiple times with different hyper-parameters and obtained a final lose and mae as 1264.5 which is also pretty good.[Jai]. Different from these examples, I want to explore a different method to build different models to achieve this goal, in order to get a higher accuracy score.

2 Exploratory Data Analysis

The following section contains several of the images I created during the exploratory data analysis.

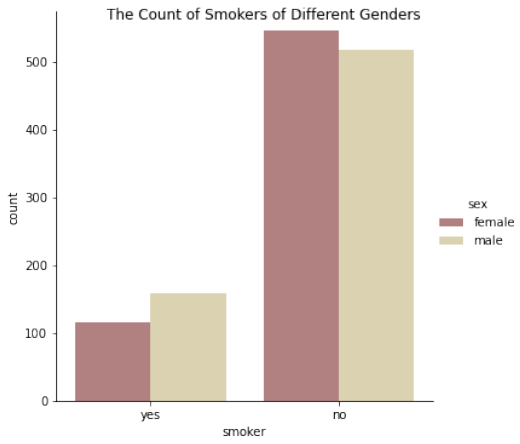

Figure 1: As we can see in the above catplot, smokers account for about 25% of the total number, and there are slightly more male smokers than female smokers.

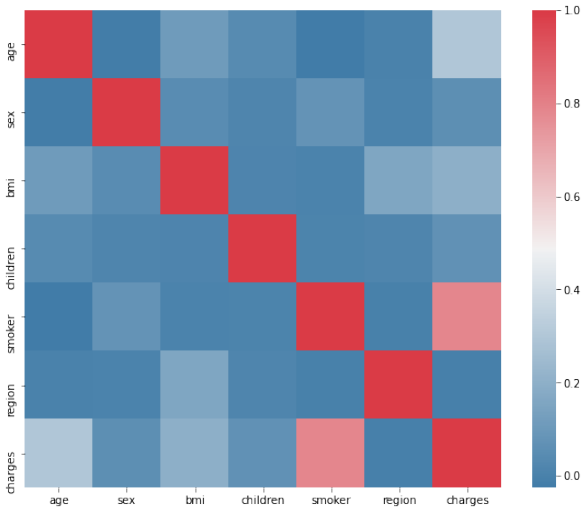

Figure 2: From the heatmap above, the influence of gender on charges seem to be negligible. Instead, bmi, age and smoking are closely related to charges, especially smoking.

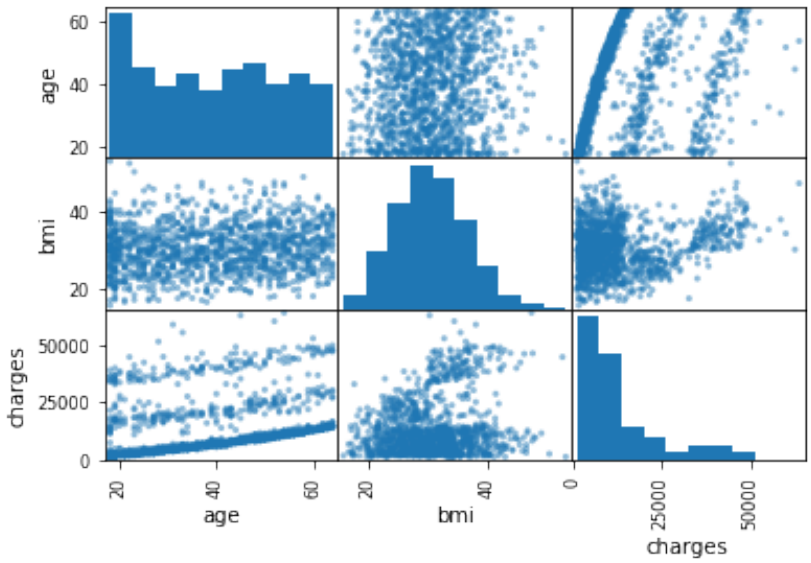

Figure 3: This scatter matrix is built from 2 features: age, bmi and our target variable charges. From these 9 plots, we can see clearly the relations between these variables.

3 Data Preprocessing

In the data preprocessing part, given each observation represents the health condition of a single individual, the data is assumed to be independent and identically distributed with no group structure or time-series property. Thus, we can just use the train-test-split to split the dataset into training data and testing data. Our dataset is not very big and larger training data would increase the accuracy of the model, so I chose to split the dataset into 80% for training and 20% for testing. In previous steps, the ordinary data sex, smoker and region data have been transferred using LabelEncoder in order to make it clear to analyse during the EDA part. While in the preprocessing part, we need to do more encoding on these features for better analyse. Since the age, bmi and children features are reasonably bounded, which are suitable for the MinMaxEncoder, and sex, smoker and region are categorical features, which are suitable for ordinalEncoder, we can use the pipeline to fit and transfer the 6 features.

References

[Jai] Praviin Jaiswal. Regression with neural network. https://www.kaggle.com/praviinjaiswal/regression-with-neural-network.

[Shr] Shubham Shrimant. Decision tree regression. https://www.kaggle.com/shubhamshrimant/decision-tree-regression.