Part 3 Report

# Physical Database Design

Physical Database is deployed on Microsoft Azure MySQL server. The code of generating the database schema is in the zip file and I added one table based on the previous ER and relation schema, which is the customer information which includes the information needed to predict customers’ expenditures and their annual charges on health areas. The attributes used in predicting is described in the following section.

# Machine Learning Design

In the previous part, I found two different files. One is from MEPS, which includes what conditions the user has and at what age they got the condition and their annual health expenditure. The other file is from Kaggle, which includes personal information such as age, number of children, whether smoking etc. This file has more attributes than the first one and thus I decided to use this file as the training data for my model.

I use only linear regression model for this dataset. I first converted the age attribute to age groups such as <20, 20-30, …, >60 and then I generated dummy variables for categorical variables such as age group, whether smoking. Then I applied the linear regression model to the dataset and used two different packages (scikit-learn, statsmodels). Both packages give out the similar high R-squared value which is around 0.752. So, I think the model is good to predict the health expenditure given those attributes.

Finally, I decided to convert the sklearn model to a onnx model for later use, because a onnx model is a better choice for inference in other applications.

# Data Architecture Design (ideas)

When the company is starting up, it probably doesn’t have much data so it can use online public dataset such as what I collected to train a model to predict the health expenditure of each customer. This method is usually not accurate due to the insufficient data in both number of records and number of attributes. And during this time, the company should build a database system that records its customers information and their expenditures just like what I designed. So that the database can store records of people and their expenditures and later when the number of customers is large, the company can use this data as the training data to train and update their models to predict customers’ expenditures. Due to the large number of records, the model is supposed to give a more accurate result based on the information the company collected when a customer is joining. So the architecture of the data storage should be database and blob storage at the beginning and then transfer to only database systems.