Tingting Wang

SEIS 763: MDL Assignment #2

1.Load the patient data from “patients.csv” file.

I was using pandas to read csv data

df = pd.read\_csv("/Users/wangtingting/Documents/763-assignment/736-a2/data/patients.csv")

2.Use variables Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus to build a linear regression model to predict the systolic blood pressure.

First, I read data into data frame, I find the variables which are numerical and categorical columns. I split numerical and categorical variables, I computed the numerical variable’s Z score and removed outlier data. Then I used One- Hot Encoding to organize the categorical variable. Then I combined all preprocessing data, finally, I was build a linear regression model.

3.What are the regression coefficients (thetas)?

Regression coefficient. It can generally be used to compare the degree of influence of independent variables on Y. The larger the Beta value, the

greater the influence of the variable on Y.

4.How do you interpret those numbers?

When age add one, the Systolic will be adding 0.5895. When the weight add one, the Systolic will become drop 1.0553, When the Gender Female is add one, the Systolic will drop 0.1971 with respired the Male. Now the best import thing is if the smoker is add one yes, the Systolic is add 10.7023.

A close up of text on a white surface

Description automatically generated

5.If you need to identify one outlier record, which record is a potential outlier? How do you reach this conclusion?

I used seaborn displot to identify outlier, I think the numerical data may be have a potential outlier, like height, weight, and age.

6.If you need to identify one or few useless features (independent variables or predictors), which one(s) will you choose? Why do you reach this conclusion?

I choose 'Location', 'Height', 'SelfAssessedHealthStatus'. Because these features have larger p-value, when I remove these features, I got the r squared number was reduced. So I think my choose is correct.