Insurance charge estimation based on regression analysis

Dake Li[1] and Likun Lin[2]

[1]Department of Mathematics, Fudan University
[2]Department of Mathematics, City University of Hong Kong

December 1, 2023

Abstract

In this article, we selected a set of data from Kaggle website about medical insurance charge, provided with BMI, age, whether smoking or not. By analysing this dataset, we aimed to predict the corresponding insurance charges for new BMI-age-smoking data. We mainly set up a linear regression model. However, in this process, we found that there was a significant difference between the data respect to smokers and non-smokers in the sample. Therefore, we classified the smokers and non-smokers into two categories and established prediction models separately.

1 Motivation and Background

The role of medical insurance in society is multifaceted. In terms of individuals, we want to build up a suitable expectation of the expenses on ourselves' medical insurance. Therefore, given personal information, a prediction model of the insurance charge is very useful. As taught in this course, linear regression model is a very powerful tool for people to estimate the relation between two variables. Since every function has a best linear approximation with its first-order derivatives in a neighbourhood, linear regression is especially useful to model some relations locally. In this article, we mainly built up a linear regression model to predict one's medical insurance charge in the future, based on his or her BMI, age, and whether smoking or not.

2 Objectives

The main objective of this article is to set up a prediction model of medical insurance charges, based on one's BMI, age, and whether smoking.

3 Data Collections and Data Preprocessing

3.1 Data Collection

We fetch the dataset from the website Kaggle (Data source: https://www.kaggle.com/datasets/mirichoi0218/insurance). This dataset contains a bunch of samples of the medical insurance charges, together with the customers' personal information. The original data has too many features, such as the customer's sex, region, have children yet or not and so on, but in order to have clearer insights, we want to take only a few features into consideration.

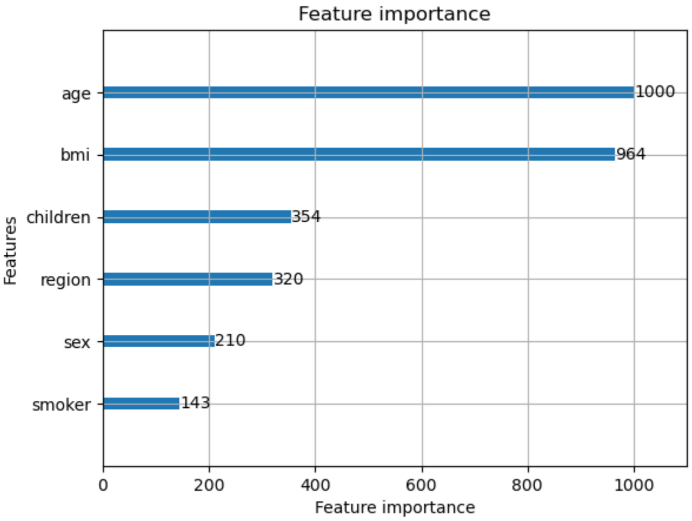
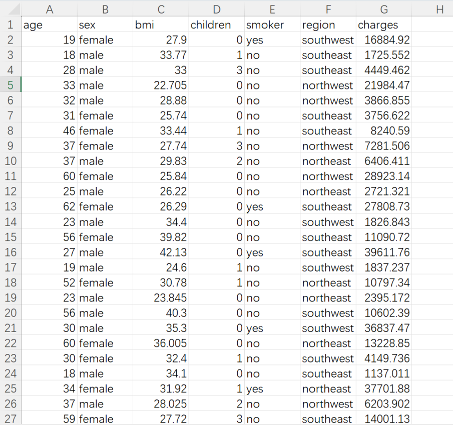


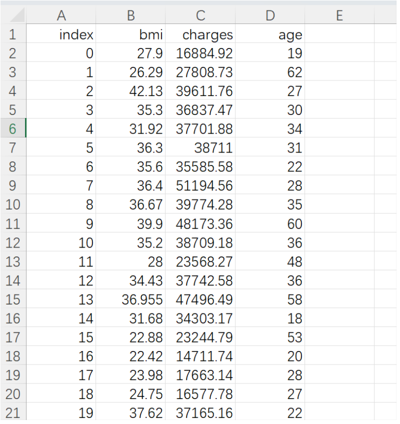
Figure 1: Feature Importance (given by xgboost)

By XGBoost analysis, we conclude that the most improtant two features are age and BMI. In addition, from our life experience, whether smoking or not is also non-negligible. Therefore, we still label our selected data with "smoker" or "non-smoker".



	A	B	C	D	E	F	G	H
1	age	sex	bmi	children	smoker	region	charges	
2	19	female	27.9	0	yes	southwest	16884.92	
3	18	male	33.77	1	no	southeast	1725.552	
4	28	male	33	3	no	southeast	4449.462	
5	33	male	22.705	0	no	northwest	21984.47	
6	32	male	28.88	0	no	northwest	3866.855	
7	31	female	25.74	0	no	southeast	3756.622	
8	46	female	33.44	1	no	southeast	8240.59	
9	37	female	27.74	3	no	northwest	7281.506	
10	37	male	29.83	2	no	northeast	6406.411	
11	60	female	25.84	0	no	northwest	28923.14	
12	25	male	26.22	0	no	northeast	2721.321	
13	62	female	26.29	0	yes	southeast	27808.73	
14	23	male	34.4	0	no	southwest	1826.843	
15	56	female	39.82	0	no	southeast	11090.72	
16	27	male	42.13	0	yes	southeast	39611.76	
17	19	male	24.6	1	no	southwest	1837.237	
18	52	female	30.78	1	no	northeast	10797.34	
19	23	male	23.845	0	no	northeast	2395.172	
20	56	male	40.3	0	no	southwest	10602.39	
21	30	male	35.3	0	yes	southwest	36837.47	
22	60	female	36.005	0	no	northeast	13228.85	
23	30	female	32.4	1	no	southwest	4149.736	
24	18	male	34.1	0	no	southeast	1137.011	
25	34	female	31.92	1	yes	northeast	37701.88	
26	37	male	28.025	2	no	northwest	6203.902	
27	59	female	27.72	3	no	southeast	14001.13	

Figure 2: Part of Original Data



	A	B	C	D	E
1	index	bmi	charges	age	
2	0	27.9	16884.92	19	
3	1	26.29	27808.73	62	
4	2	42.13	39611.76	27	
5	3	35.3	36837.47	30	
6	4	31.92	37701.88	34	
7	5	36.3	38711	31	
8	6	35.6	35585.58	22	
9	7	36.4	51194.56	28	
10	8	36.67	39774.28	35	
11	9	39.9	48173.36	60	
12	10	35.2	38709.18	36	
13	11	28	23568.27	48	
14	12	34.43	37742.58	36	
15	13	36.955	47496.49	58	
16	14	31.68	34303.17	18	
17	15	22.88	23244.79	53	
18	16	22.42	14711.74	20	
19	17	23.98	17663.14	28	
20	18	24.75	16577.78	27	
21	19	37.62	37165.16	22	

Figure 3: Part of Selected Data

3.2 Data Preprocessing

Firstly, it is observed that the data label of ’smoker’ and ’non-smoker’ led to a significant gap in the insurance charge, just as figure 1 and figure 2 show.

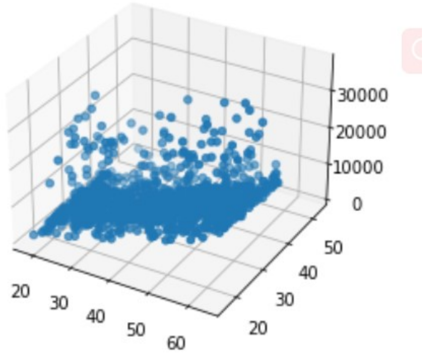


Figure 4: Nonsmoker’s condition

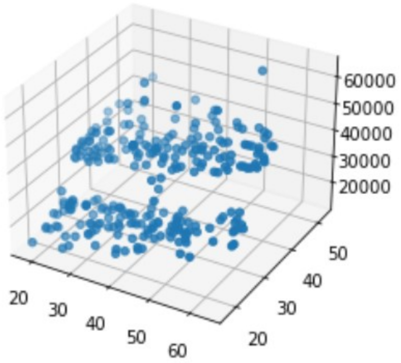


Figure 5: Smoker’s condition

According to this observation, we intend to develop different prediction models for each of the condition.

For each category(smoker or non-smoker), we randomly separate the raw data into two parts, that is, 70% of the raw data to be the training set, and the remaining 30% to be the test set.

4 Clustering and Classification

4.1 Smokers’ condition

Firstly, we begin with the category of smokers.

By observation of the BMI-charge(figure 3) and the age-charge(figure 4) figure, it is reasonable to believe that the category of smokers can be also divided into two parts, where each part possesses different properties.

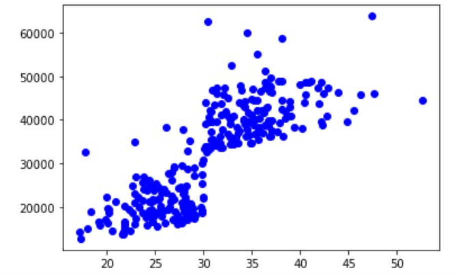


Figure 6: BMI-charges figure

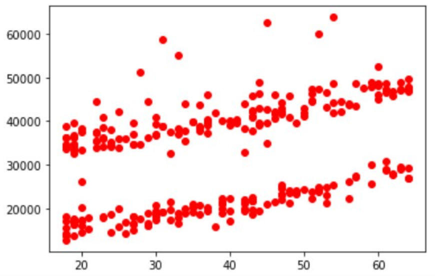


Figure 7: Age-charge figure

The following pair-plot figure provides more evidence that the dataset of smoker category has some intrinsic partition.

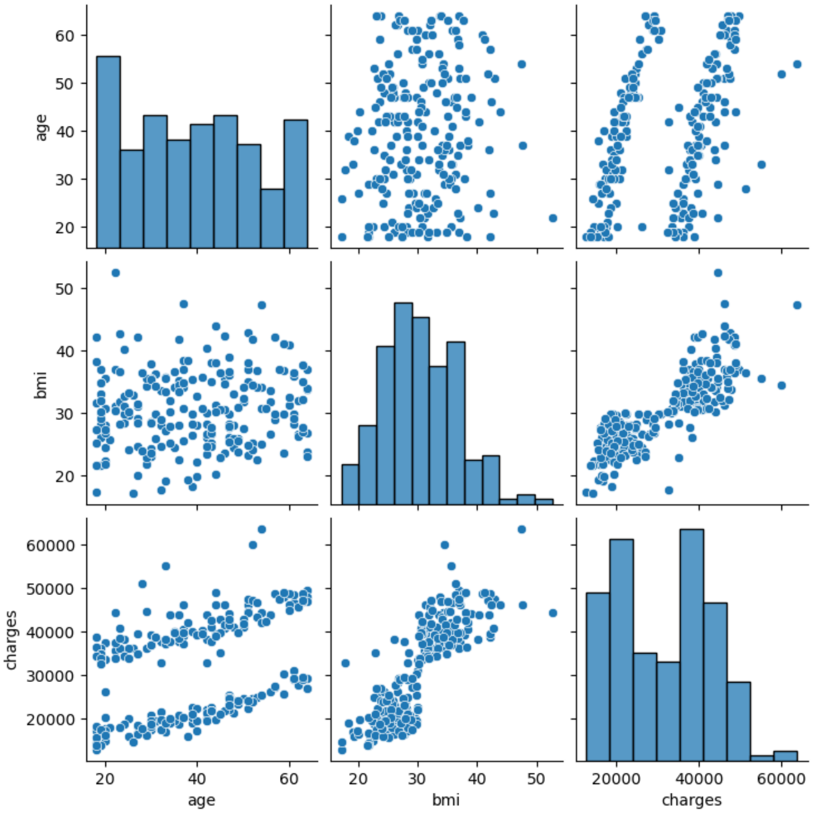


Figure 8: Pair-plot: age, BMI, charges

We conclude that the category of smoker can be further divided into two parts. The next idea is to apply linear regression on each part one by one.

The specific steps are:

1. Do clustering on the training set(70% of the whole smoker data).
2. Label the clusters by "class 1" and "class 2".
3. Using the obtained label, train a SVM model to classify.
4. Do linear regression on the each cluster.
5. For test set, we use the trained SVM to assign the label and apply the linear regression.

The following figures show the result of out clustering:

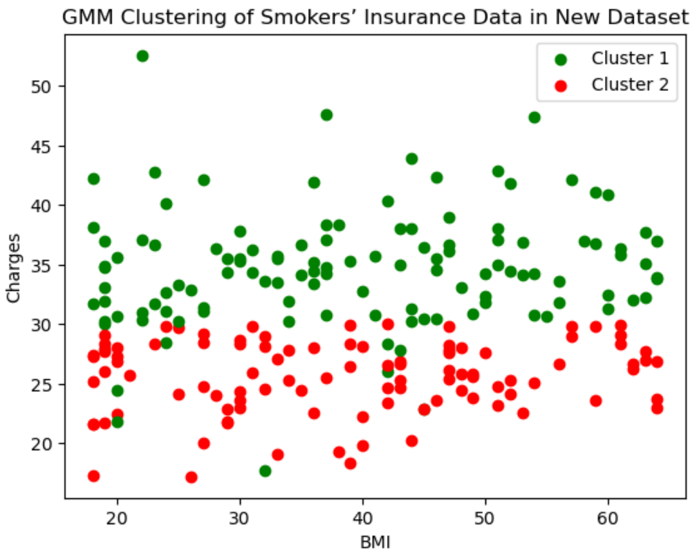


Figure 9: GMM Clustering of Smokers Category

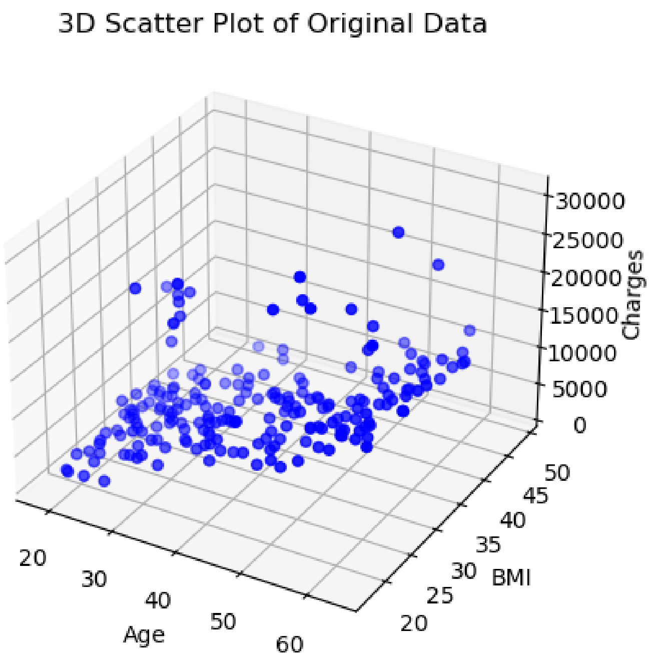
这里需要一张clustering的结果图片，要画成3维的

Figure 10: 3D plot for clustering result of smokers

According to the above analysis, the smoking population data can be divided into two clusters: high charge and low charge.

4.2 Non-smokers' condition

By observation of the scatter plot, we can still find two clusters in the non-smoker data.



We apply the same steps shown in the smokers' condition,

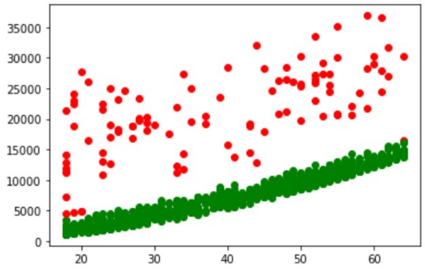


Figure 11: GMM result under nonsmoker condition

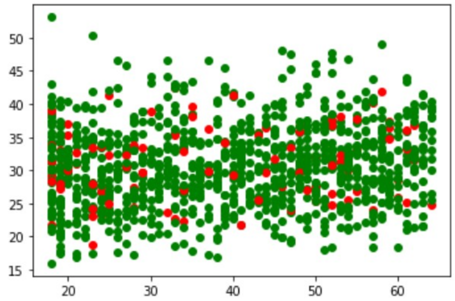


Figure 12: GMM cluster result on age bmi of non-smoker condition

5 Ordinary Linear Regression model

We use the ordinary least squares method to do the linear regression. To be more specific, suppose $(x_1^i, x_2^i, y^i) \in \mathbb{R}^2 (i = 1, 2, \ldots, N)$ are our data points, where x_1^i denotes age, x_2^i denotes BMI, and y^i denotes charge. Our goal is to find a vector $\mathbf{w} = (b_0, b_1, b_2) \in \mathbb{R}^3$ such that $\mathbf{w}$ minimizes the loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \left[y^i - (b_0 + b_1 x_1^i + b_2 x_2^i)\right]^2 = \|\mathbf{y} - X\mathbf{w}\|_2^2$$

where

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} \in \mathbb{R}^N, \quad X = \begin{pmatrix} 1 & x^1_! & x^1_2 \\ 1 & x^2_! & x^2_2 \\ \vdots & \vdots & \vdots \\ 1 & x^N_! & x^N_2 \end{pmatrix} \in \mathbb{R}^{N\times 3}.$$

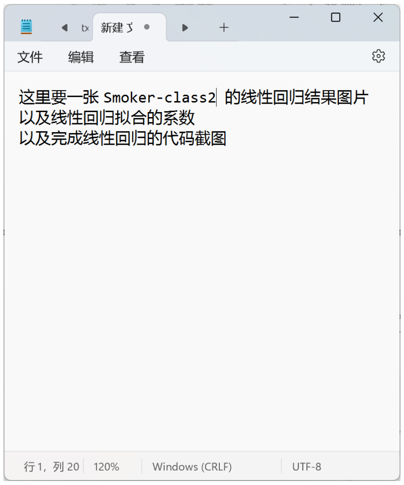
To minimize the loss function, we seek for stationary point. Let $\nabla L(\mathbf{w}) = 0$ we get

$$\mathbf{w}^* = (X^T X)^{-1} X^T \mathbf{y}.$$

The result of linear regression for smokers category:

这里要一张 Smoker-class1 的线性回归结果图片
以及线性回归拟合的系数
以及完成线性回归的代码截图

Figure 13: Smoker:class 1



这里要一张 Smoker-class2 的线性回归结果图片
以及线性回归拟合的系数
以及完成线性回归的代码截图

Figure 14: Smoker:class2

The result of linear regression for non-smokers category:

这里要一张（几张）non-smoker 的线性回归结果图片
以及线性回归拟合的系数
以及完成线性回归的代码截图

Figure 15: Non-smoker

这里要一张（几张）non-smoker 的线性回归结果图片
以及线性回归拟合的系数
以及完成线性回归的代码截图

Figure 16: Non-smoker

6 Conclusion

This paper focused on the relationship between 'charge' and its independent variables 'smoker', 'BMI', and 'age'. Our discussion is based on the fact that different

’smoker’ labels would lead to a significant difference in the value of charges. Further experiments proved that ’smokers’ can be parted into different classes for their charges with respect to ’age’ and ’BMI’, while ’non-smokers’ does not produce such a separating effect so only one class is adopted in the result. Generally speaking, the charge of the above three classes all possess a strong positive linear relationship with age. The result shows that our model may produce more errors when predicting the non-smoker category. It is suggested that relevant data, e.g. family status, is lacking in the present model, and thus further research can be conducted with aims of observing the relation between family and charge.

7 Appendix

Data source:
https://www.kaggle.com/datasets/mirichoi0218/insurance
.csv version:
https://drive.google.com/file/d/1_VQk7_KoigrBS3NEPzBX5rdbI6CE4w43/view?usp=sharing
Python Prepossessing code:
See attachment