Insurance charge prediction model

Dake Li[1] and Likun Lin[2]

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

December 4, 2023

Abstract

In this article, we explore a dataset from Kaggle concerning medical insurance charges, which includes variables such as Body Mass Index (BMI), age, and smoking status. Our objective is to predict insurance charges for new entries based on BMI, age, and smoking habits. To achieve this, we initially employed a linear regression model. However, our analysis revealed a notable disparity in the data when comparing smokers and non-smokers. This observation led us to segregate the dataset into two distinct categories based on smoking status, allowing us to develop more accurate predictive models for each group.

1 Motivation and Background

The significance of medical insurance in society is multifaceted, particularly for individuals aiming to establish realistic expectations regarding their medical insurance expenses. To assist in this, a predictive model for insurance charges, based on personal information, proves highly beneficial. As we have learned in this course, the linear regression model is an exceptionally potent tool for estimating the relationship between two variables. This effectiveness stems from the fact that every function can be closely approximated by its first-order derivatives within a local area, making linear regression particularly valuable for modeling relationships in a localized context. In this article, our primary focus has been on developing a linear regression model that forecasts an individual's future medical insurance charges, taking into account factors such as Body Mass Index (BMI), age, and smoking status.

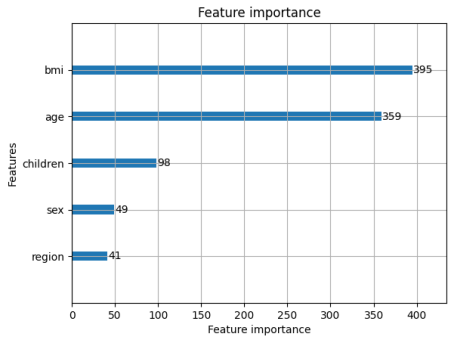
2 Objectives

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.

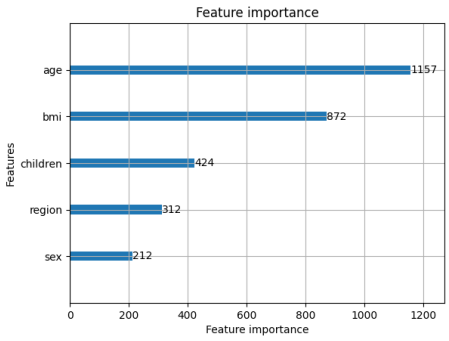
3 Data Collections and Data Preprocessing

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.



(a) Smoker's condition



(b) Non-Smoker's condition

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
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.

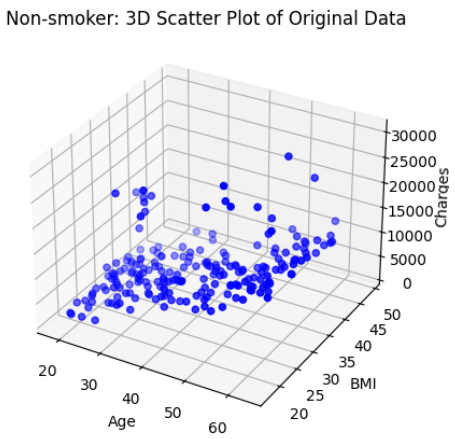


Figure 4: Nonsmoker's condition

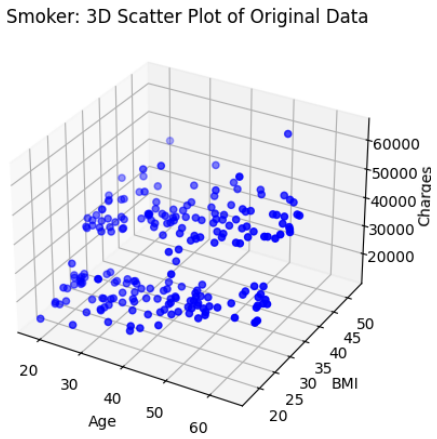


Figure 5: Smoker's condition

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 Smokers' condition

4.1 Clustering and Classification

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.

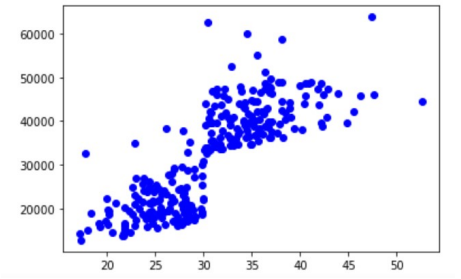


Figure 6: BMI-charges figure

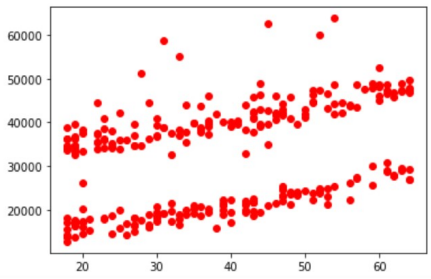


Figure 7: Age-charge figure

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

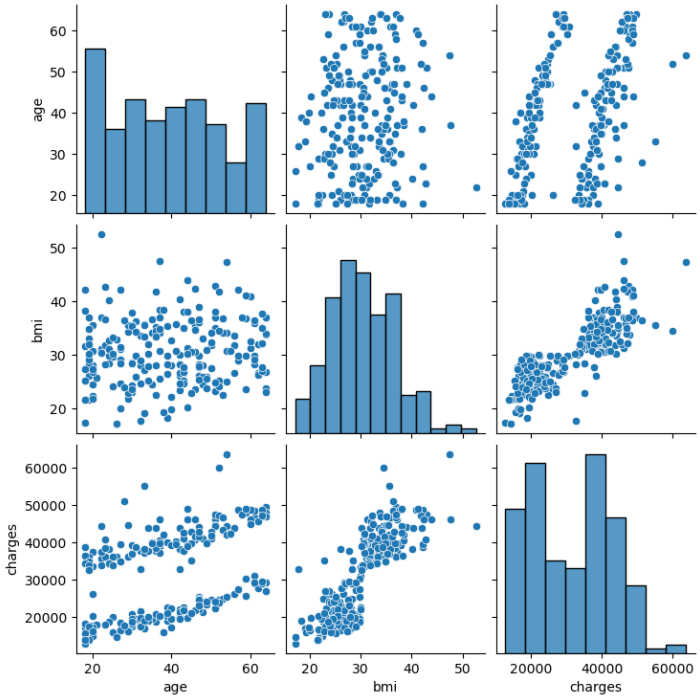


Figure 8: Pair-plot for smokers: age, BMI, charges

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 clustering:

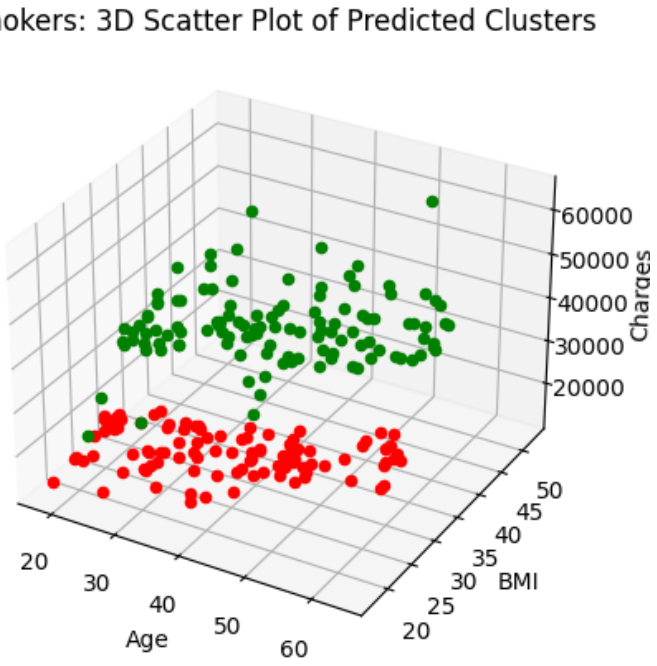


Figure 9: 3D plot for clustering result of smokers

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

4.2 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

$$\mathbf{y} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} \in \mathbb{R}^N, \quad X = \begin{pmatrix} 1 & x_!^1 & x_2^1 \\ 1 & x_!^2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_!^N & x_2^N \end{pmatrix} \in \mathbb{R}^{N \times 3}.$$

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

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

4.3 Results

The result of linear regression for smokers category:

```
class1.coefficients
[281.34877777 598.31222556]
class1.r2_score
0.7358833028942893
```

Figure 10: Smoker:class 1

```
class2.coefficients
[271.03754106 445.33344875]
class2.r2_score
0.9591511439700683
```

Figure 11: Smoker:class2

5 Non-smokers' condition

We want to apply the same steps shown in the smokers' condition, however, this method seems not to work well.

We turn to apply XGBoost to develop the predictive medel for non-smokers category. XGBoost (eXtreme Gradient Boosting) is an efficient, flexible, and scalable gradient boosting library. It's an ensemble learning algorithm based on CART (Classification and Regression Trees) decision trees.

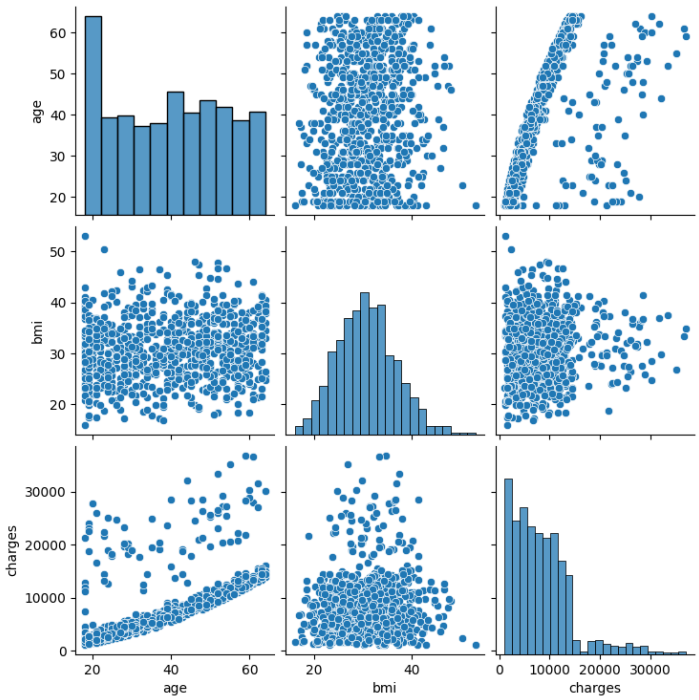


Figure 12: Pair-plot for non-smokers: age, BMI, charges

The results for this part is shown in the next section.

6 Results comparison

We compared the different results of different data processing methods, and adopted R2 scores to indicate the quality of our linear prediction model. The following table shows the R2 scores of different results:

R2 scores for smokers

	train	test
with clustering	0.957363	0.885283
without clustering	0.766306	0.70184

R2 scores for non-smokers

	train	test
XGBoost	0.436465	0.464572
ordinary linear regression	0.38314	0.400629

7 Conclusion

This paper focuses 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 indicated that 'smokers' can be divided into different classes for their charges with respect to 'age' and 'BMI', while 'non-smokers' present a much more complex behaviour. Generally speaking, the charge of the above three classes all possess a strong positive linear relationship with age.

The results show that our model may produce more errors when predicting the non-smoker category. It is suggested that relevant data, such as family status, is lacking in the present model, and thus further research can be conducted with the aim of observing the relation between family and charge.

8 Appendix

Data source:
https://www.kaggle.com/datasets/mirichoi0218/insurance
Python code and additional supporting materials:
https://github.com/fire-lit-band/MATH3320insurance