Big Data Technologies

COURSE WORK 1

SHAN KUANG (REGNO: 202381266)

Table of Content

Chapter 1 - Introduction

Having healthcare insurance is an essential part of our lives, as it allows us to be worry-free about medical expenses in most circumstances. Many countries do not provide public healthcare services, so the need for private insurance arises. Private health insurance does offer several advantages over public ones. Moreover, private health insurance can provide a number of benefits, including a broader coverage of diseases, higher reimbursements, a shorter waiting period, and others.

Approximately 8.6% of 326 million U.S. citizens were uninsured in 2021; the total number of people insured by private health insurance is 68.4% according to research (Rosso, 2023). In spite of the fact that the NHS provides free health insurance to all permanent residents (NHS entitlements: Migrant Health Guide 2014), 10.4% of them have purchased private health insurance due to a variety of reasons (Vankar, 2023).

In spite of the benefits healthcare insurance provides, the cost of coverage is high and varies from person to person. How does the insurance companies calculate your cost? How much would I spend on healthcare claim and does it really worth the money for me? Many people are wondering about this kind of problem when they choose this extra health security service. As a result, by understanding the factors that will influence healthcare costs, it will be easier for us to ascertain the cost be billed for the insurance expense.

Chapter 2 - Dataset General Analysis

2.1 Aim

The aim for analyzing this dataset is to understand the factors relate to healthcare claims amount and to predict the future cost in new policyholders.

2.2 Source

This healthcare insurance data set is freely accessible on the Kaggle website. This data includes 1338 rows for policyholders, and 7 attributes describing their identity information including:

- Age: The policyholder's age.
- Gender: Male or female of insured.
- BMI: Body Mass Index is a measurement based on weight and height.
- Children: The number of dependents.
- Smoke: The Smoking statues of the person.
- Region: The geographic area of coverage.
- Charges: The claims happened on each individual. Let's assume it's in dollar.

2.3 Analysis

2.3.1 Descriptive Analysis

The descriptive analysis can help us understand relationship between each attribute and help decide where to further explore.

Descriptive Analysis based on Numeric Columns

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

According to Figure1, we can understand the basic information about numerical values of this data set.

Figure 1 Descriptive Analysis based on Numeric Columns

- **Age:** The age range of this healthcare insurance is 18 to 64 years old. The average policyholder age is 39 years old, and half of the policyholders are under 39 years old.

- **BMI:** The BMI value ranges from 15.96 to 53.13. The average BMI in this dataset is approximately 30.66, which by the BMI standard indicates that most individuals

are obese.

- **Children:** The number of children ranges from 0 to 5, with half having one or fewer.

- **Charges:** This sector has the most significant difference range from $1121.87 to $63770.42. The average insurance claims is $13270.42 which is higher than the median, suggesting that many of high claims may increase the average.

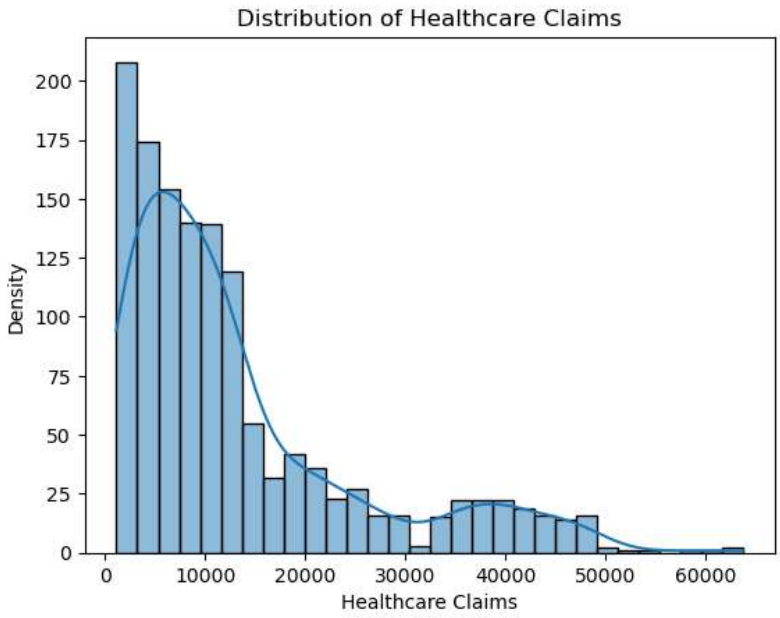


Figure 2 Histogram of Charges

A histogram has been applied (Figure 2) to further understand the distribution of healthcare claims amount. This figure shows a right-skewed distribution with a long tail on the right. Demonstrate that although the majority of patient has a claim around $15,000 and there are fewer patients have uncommon high claims.

Descriptive Analysis based on Categorical Columns

Summarizing the information based on the category column for sex, smoking condition and region, co-relation of factor and claims amount can be found. See Figure 3:

- Male policyholders claim more than Female.
- Smoker claims is 3.8 times higher than Non-Smoker.
- Among 4 regions, Southeast have the highest claim average.

Variable: sex

	Number of Policyholders	Average Claim Amount
Male	676	$13,956.75
Female	662	$12,569.58

Variable: smoker

	Number of Policyholders	Average Claim Amount
Non-Smoker	1064	$8,434.27
Smoker	274	$32,050.23

Variable: region

	Number of Policyholders	Average Claim Amount
Southeast	364	$14,735.41
Northwest	325	$12,417.58
Southwest	325	$12,346.94
Northeast	324	$13,406.38

Figure 3 Descriptive Analysis Based on Category Value

2.3.2 Correlation Between Factors

Description data has been changed to numerical, same for below chapters:

- Sex: male=1, female=0
- Smoker: yes =1, no =0
- Region: southeast =0, northwest = 1, southwest = 2, northeast = 3

A heat map between factors has been applied to explore further the distribution of individual features (Figure 4).

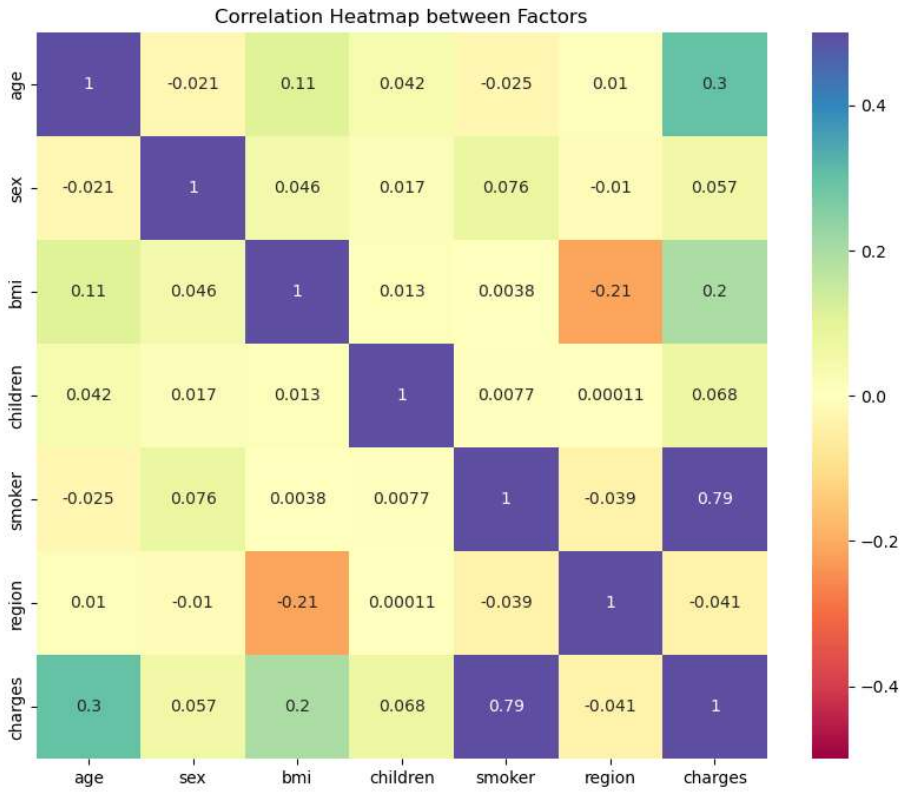


Figure 4 Influence Factors Heatmap

In view of the fact that many values are close to 0, such as the group age and gender, age and children, or BMI and children, these correlation groups indicate weak or negative correlation. Value dimensions have been set from -0.5(min) to 0.5(max) to enhance the viewing of the results.

Additionally, to the strong positive correlation (around 0.79) between smoking status and charges expected from descriptive analysis, the group of BMI and charges (around 0.2) and the group of age and charges (around 0.3) also demonstrated moderate positive correlations. There is evidence that with age and BMI increasing, the insurance claim also tends to increase, where age is more significant than BMI.

The other groups related to charges have a very weak correlation. Gender (0.057) and number of children (0.068) affect the claim amount in a slight positive relationship, but they are not statistically significant. Whereas region has a slight negative affect.

There needs to be a deeper investigation into those weak relationships.

Besides all of those factors relative to the charges, there are some interesting findings through the heatmap. For instance, there is a moderate negative correlation (-0.21) between BMI and region, this may suggest that the BMI values in certain region are lower compared than others. Many others including age and BMI, gender and smoking, age and children can be further discussed.

2.3.3 Grouped Analysis with Healthcare Claims

Thanks to the heatmap, we have gained a better understanding of the relationships between groups of factors. As a result, age, gender, BMI, and the number of children all contribute to the increase in healthcare claims.

Healthcare Charges and Children Number

Figure 5 illustrates the relationship between the number of dependents and the amount of claims. Compared to people with 1-2 children, people with 0 children have slightly higher claims on median. While the number of dependents climbs from 1 to 4, the median claim average increases. Children with five children have a lower claim amount than those who do not have children. It is possible that this is because very few people have four or more children, or they have extra ability to purchase healthcare insurance, so the sample size is small.

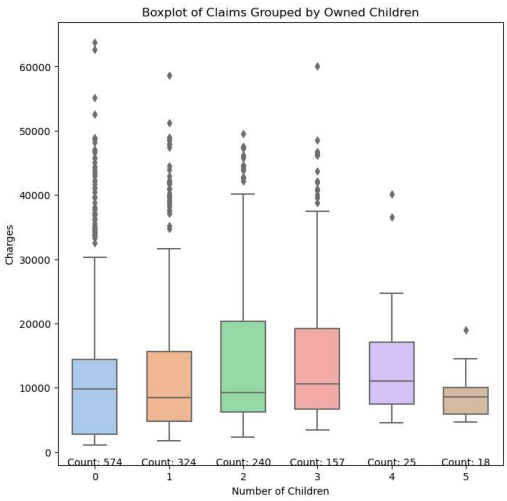


Figure 5 Charges and Children Number

Healthcare Charges and Age

With age increase, the average claims also continue to rise. Figure 6 illustrates a box plot of age and charges. For ease of comparison, we have divided our sample age range into 18-28, 28-38, 38-48, 48-58 and 58-64. It can be seen from the graph that the Q1 increases steadily across groups, while the median exhibits the same trend. The upper whiskers, however, do not demonstrate any significant trend. When comparing the IQRs of the boxes, 48-58 and 58-64 have shorter IQRs, indicating that those two groups have more consistent claims. Besides the IQRs, the groups for 28-38, 38-48, 48-58 have longer tails than the other two groups, showing that some individuals have unusually high claims which are influencing the outliers.

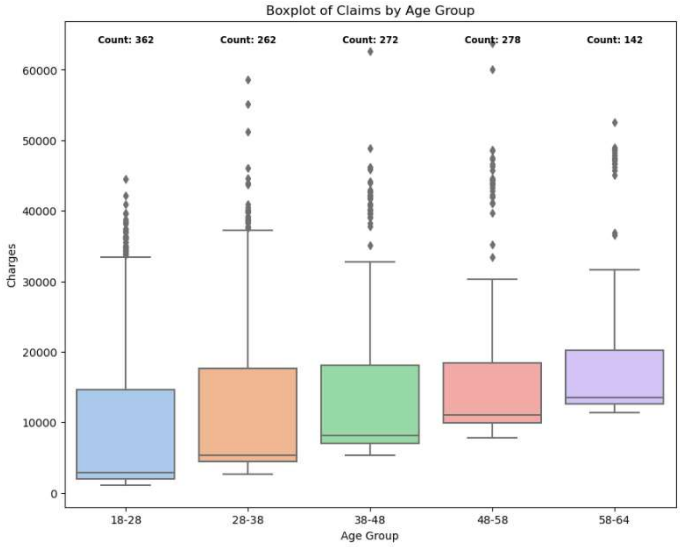


Figure 6 Charges and Age

Healthcare Charges and BMI

The World Health Organization (WHO) classifies BMI for adults into the following categories:

Underweight	Normal weight	Overweight	Obese
< 18.5	18.5–24.9	25–29.9	⩾ 30

In accordance with WHO categorization and our sample information (min 15.96, max 53.13), a bin has been set as follows: 15.9-18.5, 18.5-24.9, 24.9-30, and 30-54.2. In light of our descriptive analysis showing that more than half of all policyholders are obese, the boxplot below does not come as a surprise. Although the median of each group changes, it is only a slight increase, so there is no big impact. Noteworthy, with the increase in body mass index, the tail length of outliers grows. When people become obese, the claim amount becomes unstable, as evident from both the long tail and wider IQR.

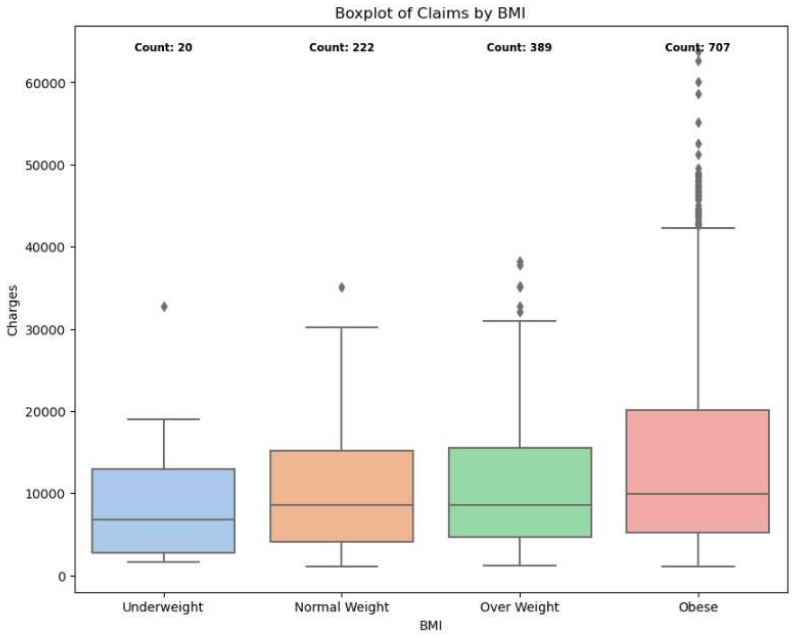


Figure 7 Charges and BMI

Considering that we have discussed the gender impact in descriptive analysis, we will not do it here since only two groups are involved. Based on the grouped analysis of claimed amounts, it appears that age has the greatest impact on the claims. A trend is only evident when 1-4 children are present in a household. As a result of the large sample size, there may be a higher median for people with no children, as evidenced by the IQR. It is also possible to apply this to the obese group based on their BMI.

Chapter 3 - Unsupervised Analysis

From above analysis and figures, we already found relationship among each group. We are going to use clustering analysis to better find the data zones and understand the methods.

The hierarchical clustering will be examined in this chapter.

3.1 Agglomerative Clustering

Agglomerative clustering is a hierarchical clustering method. In contract with Divisive Clustering, agglomerative methods start with a single group cluster and then merge gradually until all points become a single cluster or meet optimal terminal conditions.

To do that, we need to find the closest group and merge them together. Repeat this step until all points are consolidated into one cluster. The minimum distance of each cluster can be examined by linkage (Scikit Learn 2023), which contains the following four methods:

- **Single Linkage:** Min distance between two clusters.
- **Complete Linkage:** Max distance between two clusters.

- **Average Linkage:** The average distance between all pairs of points across the two clusters.
- **Ward Linkage:** The two clusters that, when merged, result in the smallest increase in variance.

Those links have their merits and demerits. Single linkage is very easy to observe, but since the required calculation is relatively simple for it, it is easy to be influenced by noise and outliers and caused chain clusters. Complete and Average linkage are able to produce clusters with relatively uniform size and shape, and they are not easily influenced, but complete linkage cannot well identify non-convex clusters than single linkage; while average has a more comprehensive calculation that requires to consider distance within each pair of points and cannot well identify closely intersecting clusters. The Ward Linkage can lead to evenly sized clusters, which can be more suitable for some data distributions since wards are trying to minimize each cluster's variance, but there needs to be a structure to the variance, so it depends on the type of data being tested.

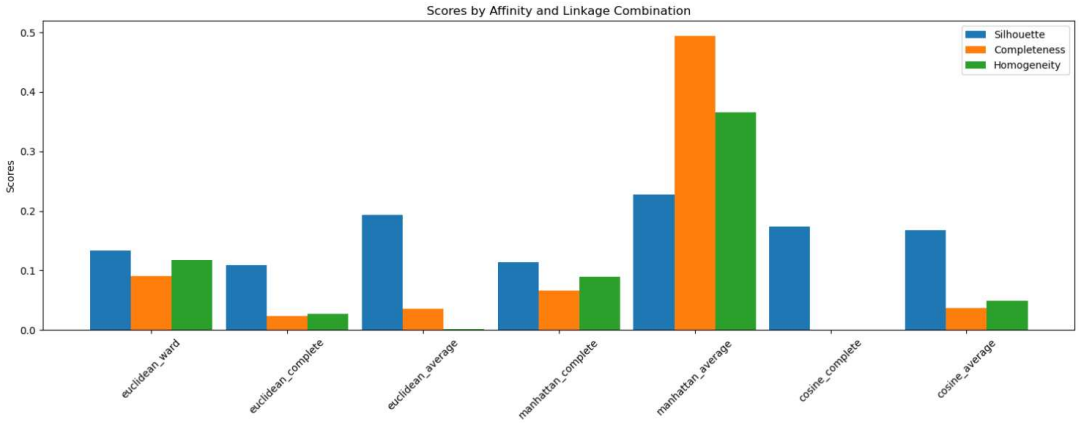
Three distance measurement methods are euclidean, manhattan and cosine similarity. Which euclidean distance is a true direct distance, manhattan is the distance of a path drawn along the axis of two points. Cosine similarity is more abstract, it's calculated as the value of cosine for the two points in a coordinate system.

Based on the above information, we tested our data and generated scores for different distance measurement methods. As shown in the graph, the used metrics are:

- Silhouette Score: A similarity of a same cluster’s point. 1 has the best quality.
- Completeness Score: If all points in the right cluster, this score will be 1.
- Homogeneity Score: If all points in the cluster are same type, this will be 1.

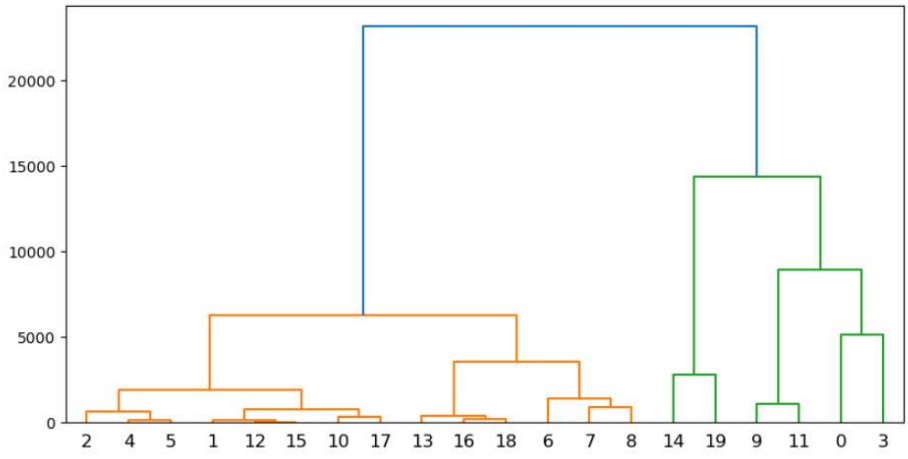
We can easily observe that among the combinations we tested, the Manhattan_Average method has the highest performance in all three scores. Euclidean_Ward, Euclidean_Average and Cosine_Average also have positive performance on Silhouette Score, means they might have high cluster quality, but poor completeness and homogeneity. In this way, the Manhattan_average has provided the best quality.

Figure 8 Performance for Agglomerative Clustering



Using Manhattan_Average, a dendrogram generate as below. If we cutting at level 5000, there are 4 clusters available. But if we cut at 15000, there will be 3 clusters.

Figure 9 Dendrogram for Manhattan_Average



Chapter 4 - Supervised Analysis

Our healthcare insurance is one of the most important parts of our lives, as mentioned in Chapter 1. Predicting the healthcare claim in a way can help us decide whether to get insurance or not. It's also the best friend for the insurance company to predict how much they should charge us. The supervised approach is an important method for machine learning. We use labeled training data - which contains paired input and output, to get regression models; and use these trained models to predict the new and unknown areas.

4.1 Linear regression

Linear regression is a method which can build up a predictive model to find a linear relationship between the goal and other variables. Using this we can predict the charge based on people's age, BMI, or their child number.

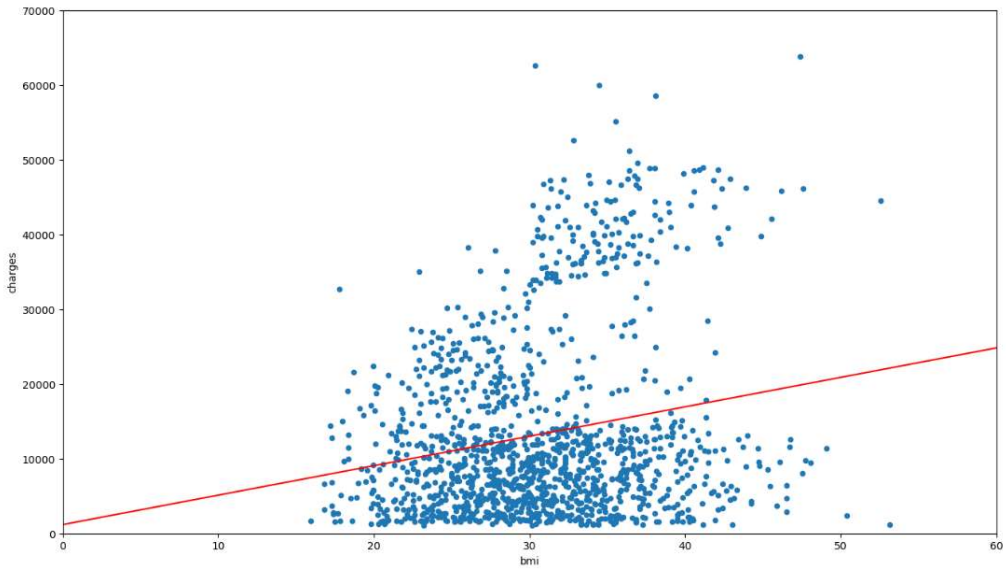


Figure 10 Trail Model for Linear Regression - Charges and BMI

We have applied the group of charges and BMI as a trail since they are continuous values. From the figure, the linear line does not perfectly fit the data. The distribution of the points tends to stick around charges under 20000, the degree of dispersion is relatively high, so our model may not give us an accurate prediction.

Using the metric below, we can see how far we have come in assessing our perfect predictions (Scikit Learn, 2023). Mean Absolute Error (MAE) is the absolute average from predict value to true value, Mean Squared Error (MSE) is the average of the square difference between predicted value and the true value, R2 is a measurement how well our model is to the prediction. Both MAE and MSE are large number, means their error of our prediction model will be big. Meanwhile, the R2 is 0.0393, which shows our model only explained 3% of the data. Linear regression is not an appropriate choice for our case.

Mean Absolute Error:	9172.351145507562
Mean Squared Error:	140777900.09850758
R^2:	0.03933913991786253

4.2 Logistic Regression

Healthcare insurance datasets have multiple categories. Logistic regression is a predictive method most suitable for solving classification problems like this. In our training model, 30% of testing data sets are total. Our goal is to understand healthcare claims. Our data has too many values; in Figure 1, there are 75% of values below 16639.1, thus the quartile is used, and two additional dividing points are added in the high value range. The charge group has been set to 6, and the bin ranges are: “1121,

4740.29, 9382.03, 16639.91, 30000, 45000, 63771" labeled "1-6".

The performance of the logistic regression is shown in the chart below. As for labels 1-3, precision, recall, and F1-scores are pretty good, especially for label 1, which has a score between 0.78 and 0.82. Labels 4 and 6 are problematic, the recall rate is 38% and 22%, leading to the model missing a lot of value in these labels and thus its performance is under-optimal. The total performance of our model is as follows, a prediction accuracy of 67% meets our expectation for this model.

Label	Precision	Recall	F1-score	Support
1.0	0.78	0.82	0.80	107
2.0	0.65	0.60	0.63	106
3.0	0.64	0.76	0.70	102
4.0	0.55	0.38	0.44	48
5.0	0.53	0.60	0.56	30
6.0	0.67	0.22	0.33	9
Accuracy	/	/	0.67	402
Macro avg	0.64	0.56	0.58	402
Weighted avg	0.66	0.67	0.66	402

In conclusion, our logistical regression model is effective at predicting charges lower

than 16639.91, but for higher charges, because the supported sample is low, our recall rate is not well performed. This may need further optimization of the model.

Chapter 5 - Reflections and Conclusion

The first course work solidified my knowledge learned from this course and reminded me of data intricacies. It has been invaluable in helping me to understand the multifaceted nature of data, and better understand its insights. Although this dataset is not too complex to analyze, the descriptive value counts for 3/7 of all variables, which limited the linear regression analysis. Also, because the subject is healthcare, the evidence from our analysis shows that there are many more variances that can influence our target - charges. I believe more information will be required to gain a deeper understanding of this subject.

From our current investigation, age, BMI and smoking habits play significant roles in our health status. But there are more relationships to investigate, for example, the region between healthcare charges: does it because of people from specific region is healthier, or it is because of the consumption level is relative lower than other regions? Those may become the questions for me until I find more data to answer.

In short, this course work is a fascinating journey for a beginner like me. The more I dig into the details, the more I find the motivation and fun in big data.

Appendix

Environment:

3.11.4 | packaged by Anaconda, Inc. | (main, Jul 5 2023, 13:38:37) [MSC v.1916 64 bit (AMD64)]

Data Source:

Kaggle https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance

Used Packages:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn
- Sys

References

1: Rosso, R.J. (2023) *U.S. health care coverage and spending - federation of American scientists*. Available at: https://sgp.fas.org/crs/misc/IF10830.pdf (Accessed: 23 October 2023).

2: NHS entitlements: Migrant Health Guide (2014) GOV.UK. Available at: https://www.gov.uk/guidance/nhs-entitlements-migrant-health-guide (Accessed: 23 October 2023).

3: Vankar, P. (2023) Population covered by Health Insurance UK 2020, Statista. Available at: https://www.statista.com/statistics/683451/population-covered-by-public-or-private-health-insurance-in-united-kingdom/ (Accessed: 23 October 2023).

4: Scikit Learn – Clustering (no date) 2.3. Clustering, Scikit Learn - Clustering. Available at: https://scikit-learn.org/stable/modules/clustering.html (Accessed: 01 November 2023).

5: Scikit Learn (no date) Sklearn.linear_model.linearregression, Scikit Learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (Accessed: 04 November 2023).