Final Project Report

EnUbi Kim, Haotian Shen, Rhiannon Abrams, Wei Liao

IST 687 – FALL 2022

Abstract: In our final project, we performed data analysis on HMO Healthcare data using data cleaning, visualization, and machine learning methods, for the suppose of making predictions on whether or not a customer will be “Expensive”. We delivered the result of our project with shiny apps website and generated actionable insights for HMO on how to lower the cost of healthcare.

According to the data pipeline, doing data science involves the following process: obtaining the data, cleaning the data, exploring, visualizing, modeling, making predictions, and interpreting the analysis. We created our project on HMO Healthcare data following the OSEMN pipeline.

1. **Data Cleaning and Exploration:**

First, we performed data cleaning on the HMO Healthcare data that had been provided to us. We had been able to get rid of some incomplete observations or used interpolation to fill the gaps.

Secondly, we inspected our data set and identified the size, as well as the potential predictors and predicted variable. There are 12 attributes here that can be used as the predictors for cost, including demographic attributes like *age, BMI, number of children, and marital status*, etc.; geographic attributes like *states* and *types of locations*; life-style related attributes like *smoker or not, exercise yearly or not*. Most importantly, after inspecting the *cost* column and learnt that the 80th percentile of cost is 5789.4, and the histogram of cost showed clearly that it has long tail effects and some people are paying way more than average. Initially, we decided to define the boundary of the *expensive* attribute to when a person paid more than $6,000.

As part of the data exploratory analysis, we visualized the data in many ways to get insights. We started by assigning *age, BMI*, *number of* *children* into categories for convenience; we had also transferred *number of children* and *education level* into binary categories (i.e., *have children or not, educated or not*) for future use. After processing the data, we created bar graphs to see if the percentage of cost being expensive differs in different groups. From those graphs, we concluded that *age*, *BMI*, and *smoker or not* might be a significant predictor since the differences in percentage among groups are obvious, while others are not.

To further explore the data, we further made bar graphs, histograms and boxplots with labels on each and every one of the variables just to get a general idea of the sample data we used. The followings are descriptions of each of the plots (also attached in R-Notebook below each chunk).

1. Age:
   1. *Histogram:*
      1. According to the above histogram, the age of most people in the data set is under 20.
      2. [Expensive - No] The distribution of this group shows a multimodal shape and a peak in the under-20 categories.
      3. [Expensive - Yes] As seen in the green area, we would say that the older a person is, the more they will pay for healthcare.
   2. *Table and bar plots: age\_group ~ number of observations*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in terms of age.

* + 1. The age group under 50 accounts for more than half of the entire population in the data set.
    2. [Expensive - No] Among the people who pay fewer costs for healthcare, the 20-29 age group has the highest proportion of the whole population.
    3. [Expensive - Yes] There are most people in the 40-49 age group with the highest healthcare cost.
  1. *Table and bar plots: age\_group ~ cost*

Unlike the above results, as we consider the cost data together, the age group between 40~59 spent a lot of money on their health care.

* + 1. [Overall] The age group 40-49 has the highest total costs and proportion in the entire population.
    2. [Expensive - Yes] Among the people with lower healthcare costs, the proportion and total value of the age group 40-49 are the highest with $4,919,178 (16.2%).
    3. [Expensive - No] The age group 50-59 pay the highest costs for healthcare services. ($3,694,369 - 12.2%)
  1. *box plot:*

The boxplots are made using the age\_group and cost columns.

* + 1. As seen in the box plot with the outliers, we can find the outliers of $55,715 on the age group over 60 with the higher healthcare cost.
    2. We can also figure out that the age group 50-59 has many outliers in the boxplot with outliers.
    3. Except outliers, the boxplots of each 'Expensive' group show similar shapes. The 'Expensive - yes' group has a more variable range of values than the 'Expensive - no' group in terms of healthcare costs.
    4. There are also outliers in the groups: the age group 20-29, 30-39, and under 18 with expensive healthcare cost.

1. BMI
   1. *Histogram*
      1. The histogram of the overall population in the BMI column shows a normal distribution (a bell shape).
      2. [Expensive - No] The distribution with a red color has a bell shape, so we would conclude this is a normal distribution.
      3. [Expensive - Yes] As seen in the green area, the shape of this histogram has a right-skewed shape. That means the data would have a higher BMI value than the 'Expensive - yes' group.
   2. *Table and bar plots: bmi\_group ~ number of observations*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in terms of BMI.

* + 1. [Overall, Expensive - No] In the data set, people who are in 'Obesity' account for most of the population. We would say these people spent fewer costs on their healthcare.
    2. [Expensive - Yes] People who are overweight may pay more costs for healthcare because the red area represents that there are 1,866 people (It accounts for 24.9% of the population)
  1. *Table and bar plots: bmi\_group ~ cost*

Unlike the above results, as we consider the cost data together, the people in the Obesity group with both low and high healthcare costs spent a lot of money on healthcare. These groups also have the highest proportion in terms of costs.

* 1. *Boxplot:*

According to the boxplots with and without outliers, the Obesity group has a wide range of healthcare cost data, and there are many outliers in the Expensive-yes category.

1. location\_type
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in location\_type categories.

In the data set, there are more people who live in urban areas in terms of both the number of observations and total healthcare costs. It accounts for almost 73-75% of the population.

* 1. *Boxplot:*

The boxplots that made by the location\_type and cost columns for the 'Expensive-Yes' group has many outliers. Considering healthcare costs, the boxplots of both country and the urban group have almost similar shapes.

1. Exercise
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in the exercise categories.

* + 1. In the data set, the number of people who exercise regularly is more than the other group without working out. Also, they have a higher healthcare cost. It accounts for almost 84-85% of the population.
    2. The interesting point is that even though there are more people who are not active and have a lower healthcare spending, the actual costs of the people who are not active and have a higher healthcare cost are higher than the other group.
  1. *Boxplot:*

There are many outliers in the not-active and expensive - yes group.

1. Smoker
   1. *Table and bar plots:*

The smoker variable has a categorical data type, so we couldn't draw a histogram.

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in the smoker categories.

* + 1. In the data set, the number of people who smoke is more than the non-smoker group. It accounts for almost 84-85% of the population.
  1. *Boxplot:*

There are many outliers in the smoker - yes and expensive - yes group. It also has a variable range of healthcare cost data.

1. Yearly-Exercise
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in the yearly\_physical categories.

* + 1. In the data set, there are more people who if the person had a well visit with their doctor during the year in terms of both the number of observations and total healthcare costs. It accounts for almost 75% of the population.
    2. The interesting point is that people who usually didn't see their doctor for a year have a higher healthcare cost.
  1. *Boxplot:*

Even though the boxplots don't have a wider range of data on healthcare costs, there are many outliers in the expensive-yes group (green boxes).

1. Gender
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) in terms of gender.

* + 1. There is no significant difference between the number of observations in female and male groups.
    2. However, in terms of healthcare cost, male has a higher healthcare cost than the female.
  1. *Boxplot:*

There is no significant difference in both female and male boxplots with healthcare costs.

1. Education level à is educated (yes or no)
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) considering whether a person has a college degree or not.

The interesting point is that people with a college degree have a higher healthcare cost than other people without an education degree.

* 1. *Boxplot:*

There are more outliers in the is\_educated - yes and expensive - yes group than is\_educated - no and expensive - yes group.

1. Married
   1. *Table and bar plots:*

The table and bar chart represents the detailed statistical results of two groups (high and low cost) considering whether a person gets married or not.

Both bar plots show that more people are married in the data set, and they have a higher cost than the other people who are not married.

* 1. *Boxplot:*
     1. As seen in the box plot with the outliers, we can find the outliers of $55,715 on the not\_married group with the higher healthcare cost.

1. Number of children à have child (yes or no)
   1. *Table and bar plots:*

The table and bar chart shows the detailed statistical results of two groups (high and low cost) considering whether a person has a child.

Both bar plots show that more people have at least one child and they have a higher cost than the other people who don't have a child.

* 1. *Boxplot:*

There are not significant differences in terms of outliers here for expensiveness between having children or not.

1. Location (states)
   1. *Map 1*: numbers of people being expensive according to states
      1. The map shows clearly that PENNSYLVANIA have more people paying more than 6000 on their health.
   2. *Map 2:* The map shows clearly that people who live on New York have higher chances of paying more than 6000 on their health.
      1. Both of the maps indicate that which state people live in might make a difference.
2. **Data Modelling and Prediction**

In the third part, we did the modelling using various techniques. We started with linear model with numeric predictors (*age, BMI, number of children*) and used cost for independent variable. We made scatterplots to see if there is a linear relationship between the three predictors and cost. The plots show that there might be positive linear relationships, but it needs more statistical evidence. The linear model established the baseline for accuracy and sensitivity. Then, we introduced supervised machine learning techniques to make more advanced predictions. The two main techniques being used here - decision trees and SVM - have the same process for dividing data into training and testing sets.

The first model we created was the decision tree model with 12 predictors from the original data, where numeric variables like age, BMI, and number of children have not been binned and put into categorical variables, and education level has not been simplified into two categories. We got a sensitivity of 0.9714 on test data set, which is actually good, but we wanted to avoid the problem of overfitting and make it more suitable for general data.

We used two approaches to make *Decision Tree Model 1* better: simplifying some of the predictors and remove insignificant predictors. For the second decision tree model, we used *age group, BMI group* and binary categories like *have children or not* and *educated or not.* Decision Tree Model 2 turned out to be better in both accuracy and sensitivity, but we were not satisfied. We then went through a selection process by eliminating some of the insignificant predictors in model 2 and included only: *age\_group, bmi\_group, smoker, location(state), exercise.* There we came up with the best performing model: Decision Tree Model 3, with sensitivity at 0.9758.

For support vector machine learning, we went through the similar process as decision tree model 1 and 2. Both SVM models underperformed compared with decision tree models.

1. **Optimization with Unsupervised Machine Learning**

Reflecting on what we had accomplished so far, we realized that we manually decided the definition of being expensive without further exploration. In the spirit of finding a more supported dividend, we used k-means clustering to put cost into groups. We exploited associate mining in the case to find the fact that BMI turned out to be most supported item associated with expensive. We included BMI along with cost to make clustering, and we chose the lowest cost in the cluster 2 (which is the cluster with second highest mean cost) as the boundary.

Later on, we repeated the process of decision tree model 2 and 3 and found our sensitivity had been improved with $12,282 as the boundary. We decided decision tree model 3 to be our model to use and stored it as “our\_model.rda” for shiny apps.

1. **Shiny Apps**

We delivered the function of making predictions with a shiny apps website. The user must upload the two datasets on the sidebars. One should have attributes including *age\_group, bmi\_group, smoker, location(state), exercise,* while the other one should contain the solution of cost being expensive or not under *expensive*. First,he or she will see the bar graphs that show the counts and percentages of people being expensive vs not expensive for different groups from our training dataset. The map tabs show the two maps we created. Clicking on the *preview tab*, they would see the data they want to make predictions on. Most importantly, by clicking on the *prediction tab*, they will be able to see the confusion matrix that encompasses the sensitivity of the predictions our model makes.

The url of our launched website is:

<https://haotianshen.shinyapps.io/FinalProj/?_ga=2.151311673.1694501232.1670083961-1568296780.1670083961>

**5. Business suggestions**

According to our discoveries, we made a few business suggestions on how to lower the healthcare cost. For starters, we found that only certain factors are needed to build a model with great sensitivity, and therefore the cost for collecting and processing the less significant data could be saved. Secondly, we discovered that certain groups of people tend to be overcharged (i.e. people with obesity, smokers or people who live in New York States). Campaigns like smoking cessation programs (targeting smokers), regular yoga sessions (targeting less active group) could be initiated to promote healthy lifestyle. Furthermore, differences among States have brought our attention to how income level, tax, and socioeconomics statues might affect health cost. The fact that people who live in New York States are more likely to be charged more might indicates that people with higher income might be charged more. In corresponding to that discovery, our suggestion is that a nationwide standard to be set for maximum cost in healthcare.