The Cost of Medical Care

Can general health factors predict insurance billing?

Corey Quackenbush
Computer Science
Hood College
Frederick, MD, USA
cq4@hood.edu

Introduction

Health coverage in the United States is critically important because unlike many other developed nations, the healthcare system in the US does not provide universal coverage for its citizens [1]. As a result, most Americans purchase health coverage through private third parties and various federal and state programs [2]. Given the current COVID-19 pandemic, it is more important than ever to ensure that people are covered by health insurance so that if they end up needing to go to the hospital, they won't get stuck paying the entire bill. This is even more critical given the fact that US medical charges are higher when compared to other developed nations [3]. This project attempts to uncover the factors that could contribute to higher medical expenses as well as examining their underlying patterns. For example, given a set of attributes, can we predict how high the associated medical bill will be? Also, can we use the data to discover unknow groups that aren't obvious from direct comparisons?

The dataset for this project was created for the book – Machine Learning with R [4] – and is available on GitHub [5]. It was created by sampling demographic statistics from the US Census Bureau. As such it approximates the real-world conditions in the United States and accurately reflects the population at large. The primary benefit of using simulated data is that there is no need to be concerned about personally identifiable information. The general conclusions drawn from this data should be able to be applied to real patient information, and still hold true, if needed. The dataset consists of 1338 entries, where each entry represents a single patient. Since we're interested in understanding general patterns that lead to higher medical costs, the attributes for each entry reflect general information about each patient. In other words, we're not interested in specific diseases – for example – that cause higher medical costs. We're interested in general factors and combinations of factors that lead to higher expenses. As such, the dataset contains the following seven attributes:

1. Age: The age of the patient. Consists of an integer ranging from 18 to 64.
2. Sex: The sex of the patient. Consists of values "male" or "female".
3. BMI: The Body Mass Index. BMI is calculated by dividing a person's weight (in kilograms) by their height (in meters squared).
4. Children: An integer indicating the number of children of the patient, i.e., the number of dependents on the insurance plan.
5. Smoker: Whether or not the patient regularly smokes tobacco. Consists of values "yes" and "no".
6. Region: The broad location where the patient lives. Consists of values "northeast", "southeast", "southwest", or "northwest".
7. Charges: The total medical expenses charged to the insurance plan for the calendar year.

Using this data, the project identifies patterns that can be used to predict future values, given known criteria. This was accomplished by exploring the data to uncover what relationships exist. This was done in a stepwise fashion, starting with a simple exploration of the data to gain a general understanding of the dataset. First, each attribute was examined, and missing data was accounted for. Then summary statistics were generated for each attribute to gain a better understanding of what they represent and what a typical value looks like. At this point, it was necessary to convert some of the factored attributes such as sex, smoker, and region, into numeric values. This depended heavily on what algorithms were run and whether or not they accepted a factored representation. It was also useful to discretize the BMI attribute to reflect the underlying use of the metric. For example, having a BMI greater than 30 means that person is considered obese. This gives us the ability to consider the groups separately. Next, the attributes were examined to uncover the relationship between them. This was done by following the correlation analysis steps in chapter 5 of Data Mining [6] and implemented with the R package – cor [7]. This will determine the degree to which the independent variables are related. After that, classification algorithms from Data Mining [6] were used to identify groups. For example, to identify unknow groups that may be present, k-means, DBSCAN, and hierarchal clustering were used to cluster the data. The main goals of the project are as follows:

1. General exploration of the data to uncover relationships.
2. Look for patterns.
3. Use patterns to predict costs.

4. Identify if there a combination of general attributes that leads to higher costs.
5. Look for hidden groups.

Methods

The data for this project was analyzed with R [8] through the IDE RStudio [9]. Many of the base packages from R were used to process and transform the data. To gain a better understanding of the data set, summary statistics were generated for each of the attributes. These values are summarized in Figure 1, below.

age	
Min.	18.0
1st Qu.	27.0
Median	39.0
Mean	39.2
3rd Qu.	51.0
Max.	64.0

charges	
Min.	1122
1st Qu.	4740
Median	9382
Mean	13270
3rd Qu.	16640
Max.	63770

bmi	
Min.	16.0
1st Qu.	26.3
Median	30.4
Mean	30.7
3rd Qu.	34.7
Max.	53.1

children	
0	574
1	324
2	240
3	157
4	25
5	18

sex	
female	662
male	676

region	
northeast	324
northwest	325
southeast	364
southwest	325

bmi_category	
Underweight	20
Normal	225
Overweight	386
Obese	707

smoker	
yes	274
no	1064

obese	
yes	707
no	631

Figure 1: Summary information about the data attributes.

Note, the attribute fields bmi_category and obese were created based on the BMI field that was present in the original data. The CDC guidelines for defining adult overweight and obesity categories were used to determine the cutoffs. Before analysis began, the data was split into a train and test set so that the validity of the models could be assessed. Approximately 90% of the data was used for training and 10% was used for testing. Next, matrix scatter plots were used to determine attribute correlation and general trends were created with the package psych [11]. Note, the Pearson correlation was used by this package to determine the correlation values in the upper right of the plot. Using the information gained from this plot, attributes correlated with medical charges were tested in a variety of regression models. In other words, the lm function within the R base [8] stats package was used to develop linear models to fit the data and to calculate summary statistics about the model. These summary statistics were used to pick the best model. After that, general data exploration was performed by creating plots through the use of the package ggplot [10]. This part of the analysis focused on the attributes that were highly correlated with the cost attribute since that field is what we're interested in predicting. Before attempting to cluster the data, the package, scales [12], was used to perform a linear conversion of attribute values in order to ensure that one attribute set didn't dominate the other. Base R packages [8] were used to perform k-means and hierarchical clustering as well as visualization of the dendrogram. Within hierarchical clustering, Ward's method was used to determine which clusters to merge. The package, DBSCAN [13], was used to create the k-nearest neighbor plot as well as cluster the data points. Elements of the package caret [14] were used to calculate the root-mean-square error and r-squared values to evaluate the regression models on the test data.

Results

The first step in the analysis was to generate a scatter plot matrix to understand the relationships between the attributes and get a sense of which ones to investigate further. Attributes with a high correlation to medical charges were then selected and used to create a linear model for regression analysis. The final attributes that were selected were then re-plotted and are shown in Figure 2, below.

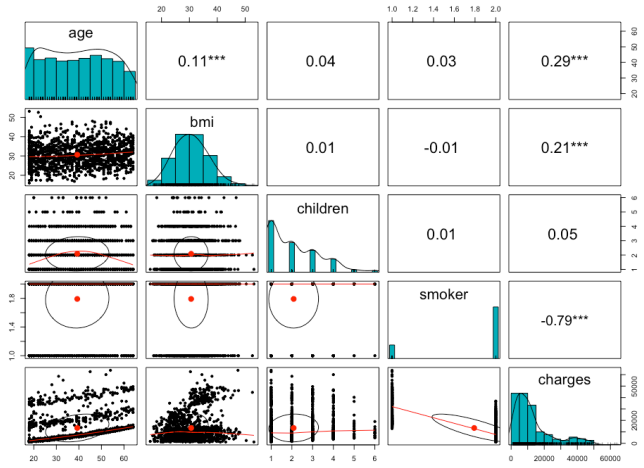

Figure 2: Scatter plot matrix - final attributes.

The regression model produced with these variables can explain about 75% of the variation seen in medical charges. Furthermore, when analyzing the model with the test set, the majority of the predictions are between about $3,000 over the true value and about $1,600 under the true value. This can be shown graphically by plotting the predicted values verses the actual values (Figure 3).

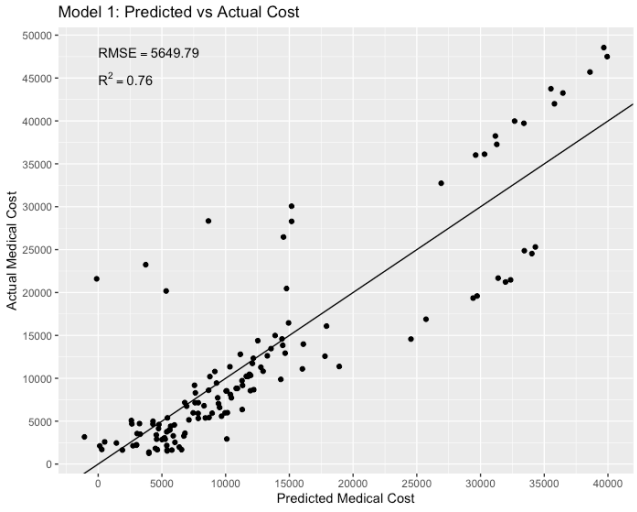

Figure 3: Regression model 1

The predictive power of this model isn't too bad but looking at the lower left side of the scatter plot matrix, there are some non-

random interactions that we might be able to take advantage of in order to improve our model. To uncover what interactions might be of use in our regression model, general data exploration was performed. The first attributes examined were charges and BMI. It can be shown, from Figure 4, that there is a second set of patients in the obese category that pay substantially more in yearly medical costs.

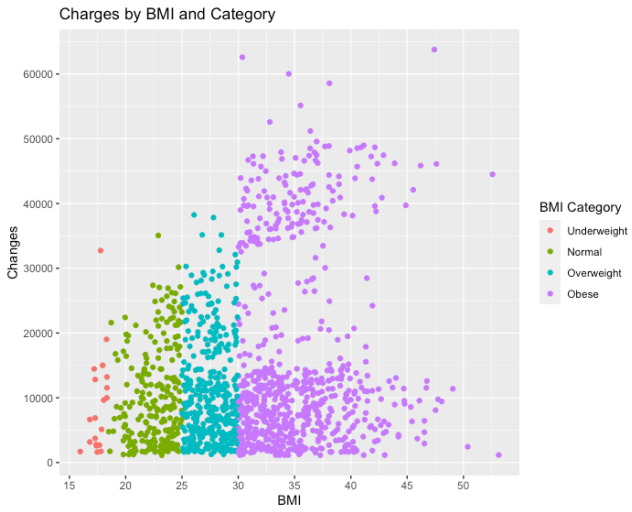

Figure 4: Identification of a second set in obese group

This clear separation of patients, resulting in higher medical expenses is a perfect example of something we could leverage to improve the predictive performance of our regression model. As such, the other attributes were compared, and another interesting pattern emerged. Figure 5 shows that there is a second set of patients in the smoking group that pay higher yearly medical expenses.

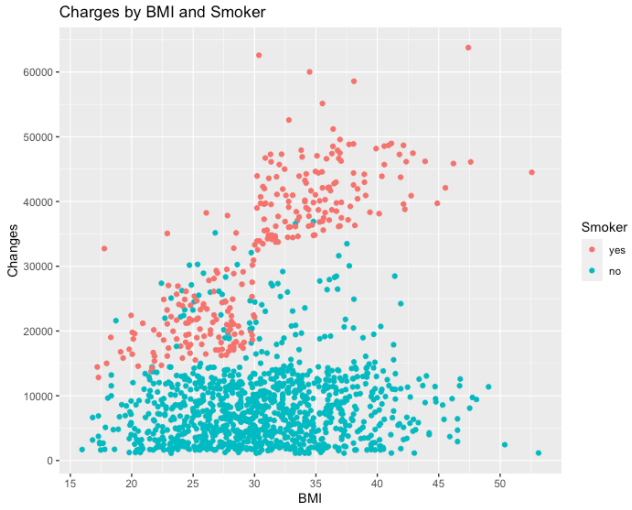

Figure 5: Identification of a second set in the smokers group

It can also be shown that this second group of individuals are obese because their BMI is above 30. To verify this relationship, the charges, obese, and smoker attributes were examined directly. The results are shown in Figure 6, below.

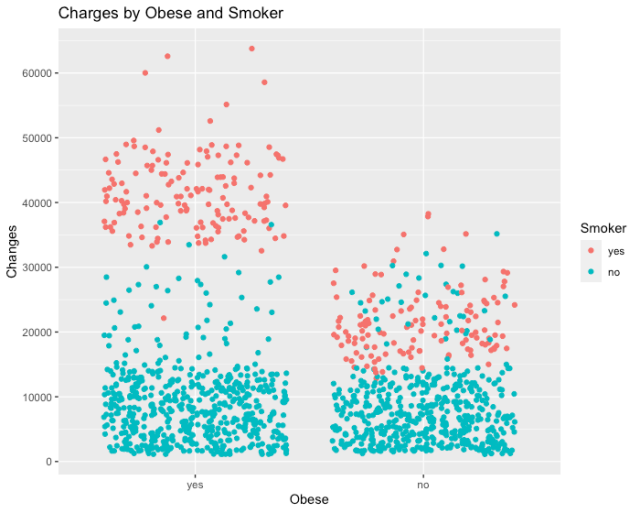

Figure 6: Identification of high yearly medical expense group

It is clear from this plot that the second group can in fact be identified by combining the smoker and obese attribute fields. We can use this information to improve our regression model by identifying that the BMI and smoker attributes have a combined effect. Testing the effect of this addition can be done by running our test set through the updated model and comparing the actual and predicted medical costs (Figure 7).

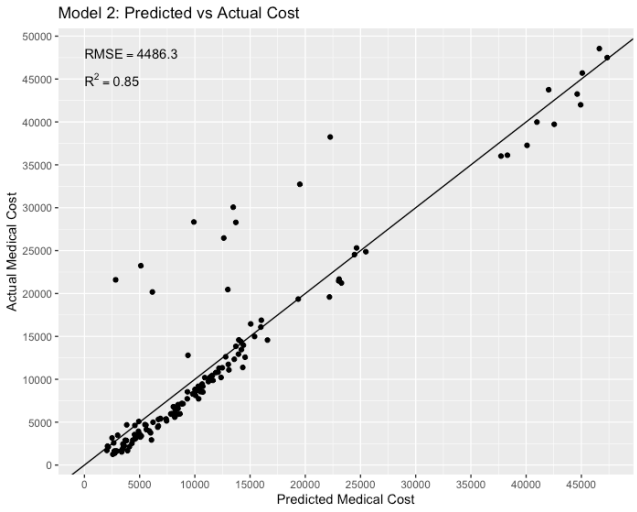

Figure 7: Improved regression model

The regression model produced with this update can explain about 86% of the variation seen in medical charges. The residuals also indicate that the majority of the predictions are now over predicted by only $400 to $1800 of the true medical expenses.

Various clustering techniques were also performed on the data to investigate whether we could capture the obese-smoker group.

The technique that performed the best was hierarchal clustering using Ward's method. The result from this clustering technique is shown in Figure 8.

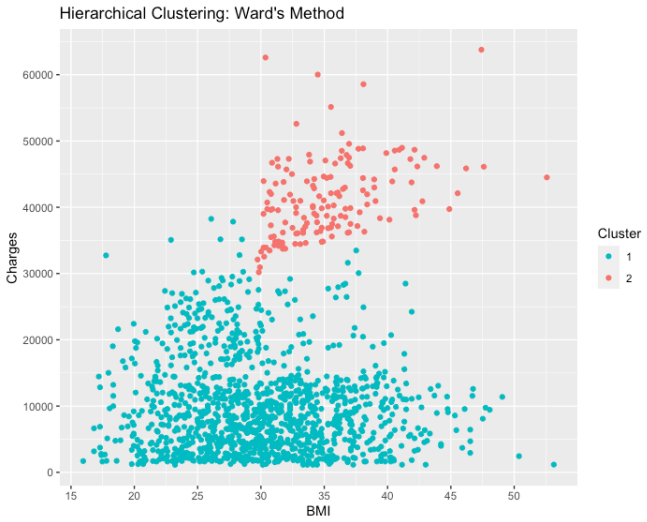

Figure 8: Capturing obese-smoker group with clustering

This method does a fantastic job of identifying the obese-smoker group with very little information. Using this method, new data can be classified by using k-nearest neighbor to determine which group the patient belongs to.

Conclusion

In summary, this project is focused on understanding what relationships exist between attributes, finding patterns in the data, and using the patterns to predict outcomes. In this case, the main outcome is being able to predict what price an individual will pay in yearly medical expenses, given a set of general factors associated with them. For example, given a patients age, BMI, number of children, and knowing if they smoke, can be used to predict their annual medical expenses. The model is able to explain 85% of the variation in medical charges and the majority of predictions are over predicted by only $400 - $1800. This information could be used to help patients budget for future expenses. For example, if the patient has a flexible spending account, they could use this information to increase or decrease their contributions to that account. This could allow them to partition their pre-tax money to pay for their medical expenses. This information could also be used to help convince a patient to become healthier. For example, a doctor could tell their patient how much money they could save by making certain life changes.

Hierarchal clustering using Ward's method was very effective at capturing the obese-smoker group. Once the data set is grouped, new samples can be categorized by either re-running the clustering algorithm or by simply calculating the k-nearest neighbor of the new points to determine which group they belong to. This method is very useful because allows us to derive information about the patient with very little information. For example, if we know the patients BMI and yearly medical expenses, we could say something about how likely they are to be a smoker. Similarly, if we know the patients BMI and whether they were a smoker, we could estimate amount of yearly medical expenses they would pay. Finally, if we knew the patient was a smoker and we knew their yearly medical charges, we could determine if they were obese or assign them to a BMI range.

REFERENCES

[1] Institute of Medicine. Committee on the Consequences of Uninsurance. 2004. Insuring America's health: principles and recommendations. Washington, DC: National Academies Press.

[2] Millman M, editor. 1993. Access to health care in America. Institute of Medicine, Committee on Monitoring Access to Personal Health Care Services. Washington: National Academies Press.

[3] Ridic G, Gleason S, Ridic O. Comparisons of health care systems in the United States, Germany and Canada. Mater Sociomed. 2012;24(2):112-120. doi:10.5455/msm.2012.24.112-120

[4] Brett Lantz. 2019. Machine Learning with R: Expert techniques for predictive modeling (3rd ed.). Packet Publishing, Birmingham, UK.

[5] Machine Learning with R: Datasets. GitHub. https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

[6] Tan P, Steinbach M, Karpatne A, Kumar V. 2019. Introduction to Data Mining (2nd ed.). Pearson Education Inc., New York, New York.

[7] RDocumnetation: Correlation, Variance And Covariance. Retrieved from: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

[8] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

[9] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com

[10] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[11] Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA.

[12] Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1.

[13] Hahsler M, Piekenbrock M and Doran D (2019). "dbscan: Fast Density-Based Clustering with R.", Journal of Statistical Software, *91*(1), pp. 1-30. doi: 10.18637/jss.v091.i01

[14] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-81.