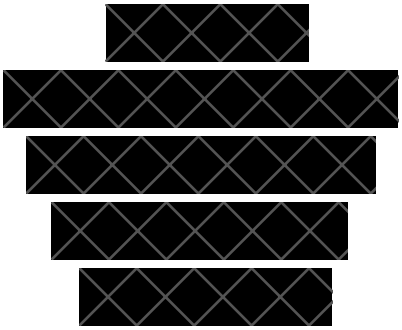
Helping Healthcare



Introduction

What makes you a risk to your medical insurer? Does having more kids mean more health concerns? How much will smoking actually cost in medical expenses? These are some of the big questions Medical insurers have to clarify. Medical insurers cover the costs of medical procedures and check-ups. They need to assess how much they should charge their clients monthly to be able to cover their medical expenses and make a profit. Therefore, the goal of this project is to be able to understand what factors attribute to a patient's medical costs. Knowing which attributes cause high medical bills, will allow the medical insurance provider to charge a higher premium.

In particular, we will focus on a dataset of patients' attributes and medical bills. Furthermore, we will focus on pinpointing which attributes cause medical expenses and we will focus on building an optimal model to accurately predict medical bills based on patients' attributes. The models utilized are linear regression, clustering, logistic regression, classification tree, and random forest.

Dataset and Exploratory Analysis

Kaggle provided the dataset. It has 9 attributes; the most important of which is "Amount Charges". Charges represent the medical cost of a patient. The other 8 attributes describe the patient's age, sex, BMI, region, number of children, and if they are a smoker or not. Please refer to **Figure 1.1** to see a complete data description. From our exploratory analysis, we were able to find that the majority of charges are under $15000 and charges is right-skewed which is exemplified in **Figure 1.2.** From **Figure 1.3**, we can conclude that there is an underlying factor(s) influencing heavy charges. Although they fall outside the expected range, they should not be treated as outliers and will be explored more in deeper analysis. We were also able to make important discoveries and inferences from **Figure 1.4**. As you age, the more likely you are to suffer higher medical expenses. Charges and BMI also had a positive relationship possibly suggesting that overweight individuals could have more intense health concerns. We also found that your BMI increases with your age.

Linear Regression

In the method of linear regression, charges is the response variable and BMI, sex, age, children, region, and smoking are the predictor variables. This model was able to uncover several significant relationships and insights. In regards to smoking, Smoking is significantly correlated with expensive health issues. Smokers cost an average of $23,847.50 more than non-smokers per year. For each additional year of age, we would expect $256.90 higher charges as age is a significant factor in determining charges. Individuals in the Northeast region incur the highest medical costs. Each additional

child costs an average of $475.5 more in charges. BMI is a significant factor contributing to medical expenses. Females experience $131 more in medical expenses. All in all, old age, smoking, and obesity are the primary factors associated with health issues. This model was able to explain 75.09% of the variability in medical expenses and uncovered relationships that will be further explored.

Clustering

In exploratory data analysis, cluster analysis groups a set of objects into groups in order to discover similarities within the data. Clustering is a form of unsupervised learning, so it does not use class. In our analysis, we performed two different clustering analysis. We begin with hierarchical clustering, so in the first attempt of clustering, the variables that we explored were the BMI, number of children, and smoking, to see if there are any similarities or possible grouping. Initially the " Amount" column of the data was saved for comparison purposes as a subset of the data. The "Amount" column is split evenly with 669 instances in both "Above Average" and in the "Below Average". The first step in hierarchical clustering is to calculate the distances, by using the distance function in R. After having the distances, the hierarchical clusters are made using the complete, average, and single methods. Complete-linkage clustering begins with each element as its own, and then each cluster is combined based on the distances until they are all in the same cluster. Next, in single-linkage clustering at each step clusters are combined based on the closest pair of elements. Lastly, the average-linkage clustering is done by computing the averages of the distances in between clusters. After plotting these three different graphs, **Figure 2.1,** it is clear that single linkage is not readable as well as the other two methods, so it was excluded from further analysis. After using the cuttree function and comparing with the "Amount" variable, we see that clustering does not produce very accurate results as the results are not split evenly as they should be. Overall, hierarchical clustering did not provide a lot of insight into our data.

In the second part of clustering, we explored K-means clustering in hopes of concluding more useful results. The first step was to determine the optimal number of k, which is how many groups of clusters should be done. To figure this out, we created a plot of clusters versus within groups' sum of squares. This can be referred to in **Figure 2.2**. Based on this graph, we proceeded with k=2. To get a visual of the clusters, we used R and plotted the data with several variables. First we looked at children versus charges, **Figure 2.3**, which showed that there is little to no similarity between the two variables. Second, we looked at smoking versus charges, **Figure 2.4**, and the data supported that a smoker is grouped with higher charges, which is supported by our other data mining algorithms. Lastly, we tested out BMI and charges, Figure 2.5, and it seems that higher charges are grouped with a larger BMI. Overall, k-means clustering helped support analysis found in other algorithms.

Logistic Regression, Odds

Logistic Regression is one of the basic algorithms to solve classification problems. Therefore, we started by updating our dataset attributes to predict the logistic regression (log) model that takes values from 0 to 1. We run our dataset on an R program to get the values of the significant coefficient based on P-results, see **Figure 3.1**.

For example, the estimate values of $\beta 3 = -0.3182$. This coefficient β3tell us when increasing children by one unit, the log odds of amount charge above average (versus below average) is expected to decrease by -0.3182, with all other predictors held fixed. Same estimate would produce coefficient $\beta 8 = -4.4405$. We run the Exp (β3) = 0.7274 and Exp (β8) = 0.01178. These exponentiated coefficients are called Odds Ratio. This means that holding sex, BMI, smoker, region and young adult at a fixed value, for a one-unit increase in children, the olds of amount charge above average (versus below average) increase by a factor of 0.7274 or -72.74% decrease in the odds of amount charged on β3, with AIC 734.73 based on our updated dataset.

To predict accuracy on our model, we split the dataset into two groups: 80% training and 20% testing. We run the R-program to predict a new logistic regression. The testing set predicts the response to our analysis with a list of 268 observations with two significant attributes (children and groups). Testing will show us 50% of the response to be above average or below average to predict our amount charged into this testing data set. This will generate our Confusion Matrix table **Figure 3.2**. This table evaluates two types of errors (False Positive or False Negative) that indicates the accuracy of the correct predictions. The table gives us only 22% accuracy meaning that this method is not efficient.

Classification Tree

A classification tree is a structural mapping of binary decisions that lead to a conclusion about the class of an object. It is composed of branches that represent attributes, while the leaves represent decisions. It's a popular method in data mining because it's easy to understand.

As we mentioned before, the classification tree is used for categorical response, so we use the dataset without the age and charges column and use Amount as our response, which includes above average and below average. Then we divide 80% of our dataset as training and the rest as testing. Our testing data contains 268 observations and 7 variables. We use k-fold cross-validation to determine the optimal subtree and apply cv.tree function to obtain the size. Based on **Figure 4.1**, size 4 and size 3 has the same lowest cross-validation error rate, which is 100 compared to 437. We decided to use 3 as our optimal size for our classification tree. **Figure 4.2** shows us the tree with three terminal nodes. From the

tree, we can see smoke is a primary factor that determines whether the amount is below or above the average. Assuming other elements are the same, people who smoke will incur a higher amount of insurance costs. Another significant factor will be the age group. People who don't smoke but are in the senior group are also more likely to have the amounts that are above the average. We define the senior group if the age is over 47. After we form the tree model, we use a confusion matrix to evaluate it by comparing it to the testing data. From the result in **Figure 4.3**, this model gives us an accuracy of 92%, which indicates a good result for this method.

Random Forest

Random forest is another method used to determine which attributes were deemed important in predicting the accuracy of class prediction. Random forest is a powerful method that reduces variance by randomly selecting the decision trees and then averaging or picking the class that gets the most votes. With this dataset, we are using classification trees in which the "leaves" represent class labels and the "branches" represent the features that lead to the class labels.

Random forest utilizes bagging which repeatedly selects a random sample with replacement of a training set to receive a prediction of the class. Random forest uses a random subset of the variables to increase prediction accuracy and prevent bias for strong predictors. For this dataset, we used an mtry parameter of 2 as it is the square root rounded down of the total number of variables considered. Each node selects from one of the two variables it has randomly chosen from the original 6 attributes based on which optimizes the split. The original 6 attributes used were sex, bmi, children, smoker, region, and (age) group. Once the model was executed, the model received an accuracy of 0.9216418 and an error of 0.07835821. This model also produced the same results for the confusion matrix as the classification tree **Figure 5.1**. After calculating accuracy, we wanted to look at the dataset's important variables according to our random forest model. Using the importance() function in R, we found the highest mean decrease accuracies and mean decrease Gini's were ones from the attributes "smoker" and "Group" - the third-highest was from the variable "bmi". The importance() function calculates the mean decrease in out-of-bag prediction error (MeanDecreaseAccuracy) and the mean decrease in the node impurities derived from splitting on the variable, in this case, from the Gini index (MeanDecreaseGini). For visualization purposes, we used the varImpPlot() function to plot these mean decreases **Figure 5.2**. The results of the importance() function tell us the top two significant predictors of the dataset are found in the columns "smoker" and "Group". It is safe to imply given the results of the classification tree and other data analysis that Yes in "smoker" and Senior Adult in "Group" are the subsets of the dataset that are significant, according to the random forest model.

Final Model and Conclusion

After analyzing all five methods, we chose the classification tree as our final model due to its easy interpretability and high accuracy. In the classification tree model, we predict 247 observations correctly from the 268 observations, which gives us 92 percent accuracy. From the classification tree analysis, our tree model will be three terminal nodes. Smoker and age group will be the critical factors that help us to decide whether the amount is above or below the average as we see in **Figure 4.2**. Smokers and senior adults will have higher insurance costs than those nonsmokers and younger adults. We highly recommend insurance companies to use our classification tree model to analyze the insurance cost. Once the average line is set, they will quickly have a sense of whether the amounts of the customers are above or below the average line based on smoke status and age group. Then the insurance company will be able to charge more on those customers who are more likely to have higher medical expenses. We believe that our model helps the insurance company operate more efficiently and effectively.

Figures

Attributes	Description
age	Age (Years)
sex	Gender (Male, Female)
bmi	Body mass index
children	Number of children
smoking	0 = Non-smoker & 1 = Smoker
region	Region of the US (Northeast, Northwest, Southeast, Southwest)
charges	Amount of a Medical Cost in Dollars($)
amount	Category of Charges (Below Average, Above Average)
group	Category of Age (Young Adult, Adult, Senior Adult)

Figure 1.1. Data Description

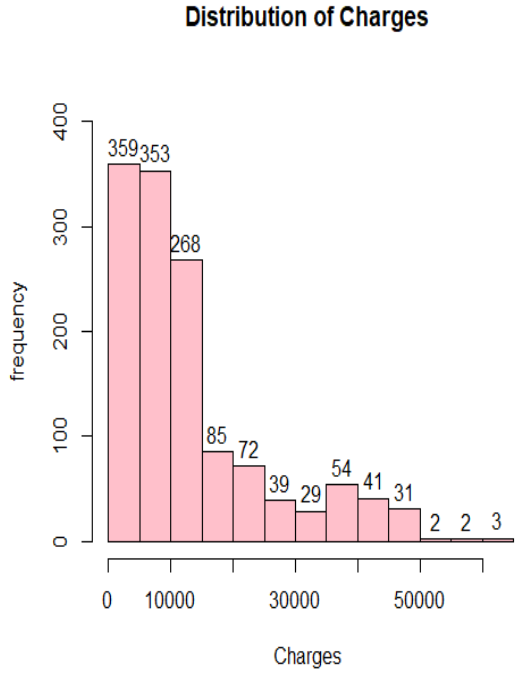


Figure 1.2. Histogram of Charges

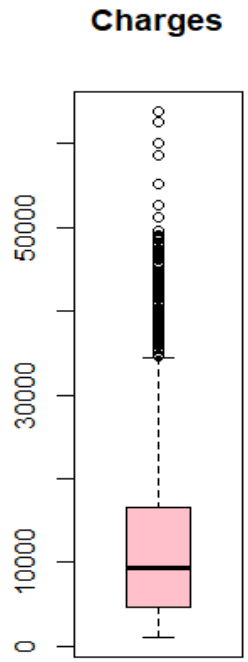


Figure 1.3. Box & Whisker plot

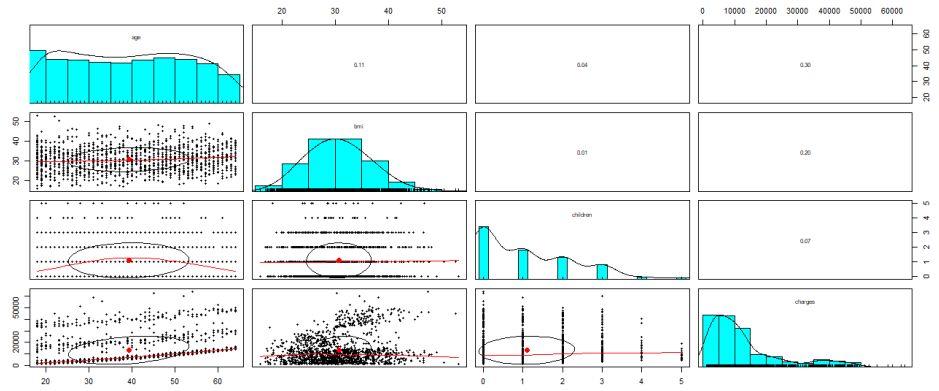


Figure 1.4. Correlation Matrix

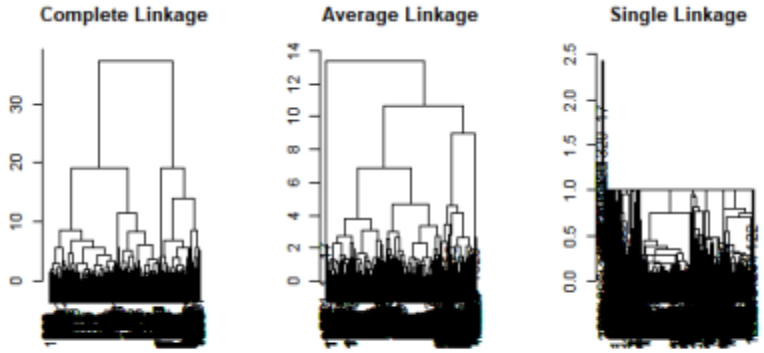


Figure 2.1 Hierarchical Clustering

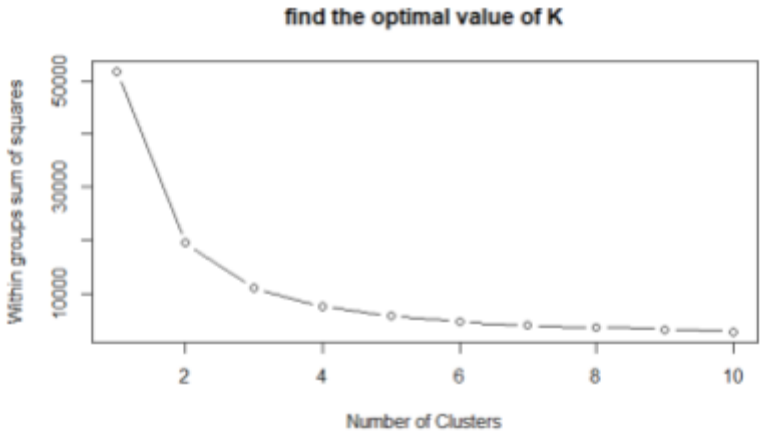


Figure 2.2 Optimal value of K chart

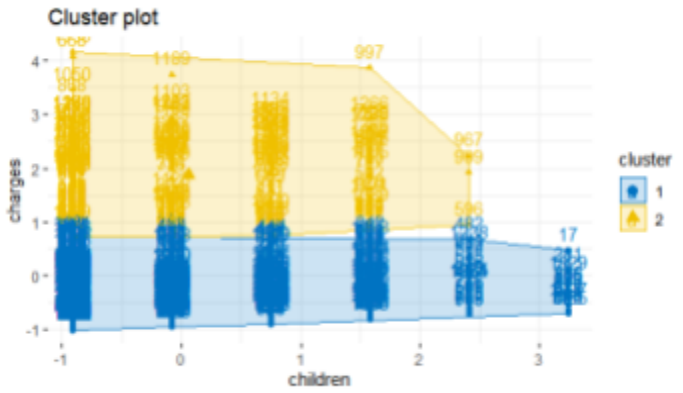


Figure 2.3 Children vs charges clustering

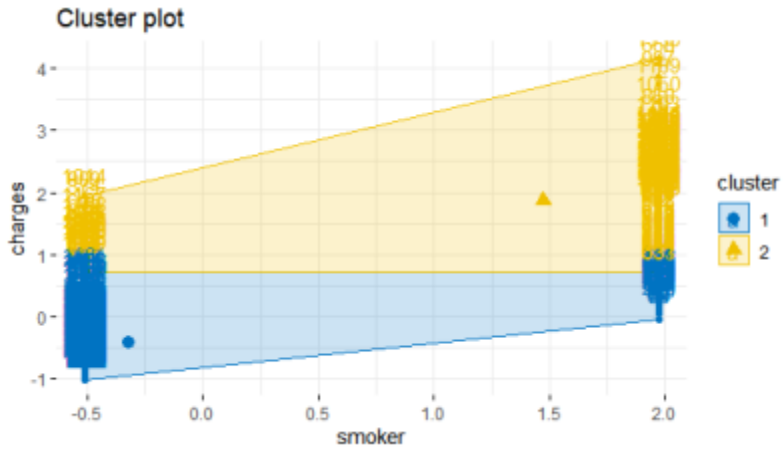


Figure 2.4 Smoker vs charges clustering

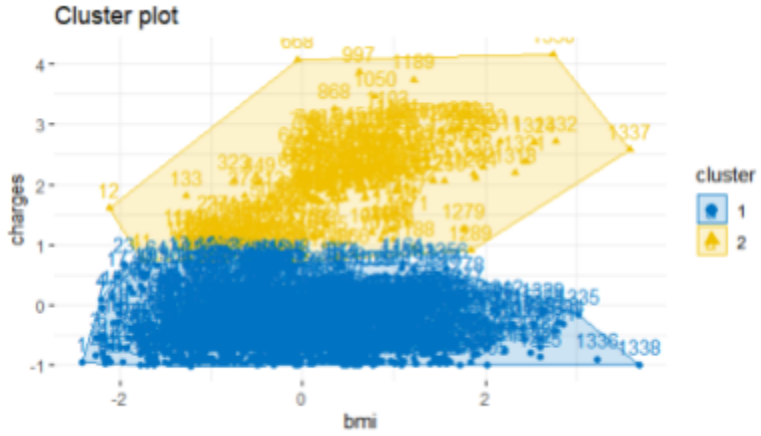


Figure 2.5 Bmi vs chargers clustering

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.30655    0.59164   5.589 2.29e-08 ***
sexmale            0.36766    0.19851   1.852 0.064012 .
bmi               -0.03783    0.01696  -2.231 0.025713 *
children          -0.31816    0.08264  -3.850 0.000118 ***
smoker           -21.29424  520.88172  -0.041 0.967391
regionnorthwest    0.44332    0.27597   1.606 0.108179
regionsoutheast    0.94717    0.29118   3.253 0.001143 **
regionsouthwest    0.86168    0.28111   3.065 0.002175 **
GroupSenior Adult -4.44054    0.26890 -16.513  < 2e-16 ***
GroupYoung Adult  -0.37401    0.27084  -1.381 0.167308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.1 Coefficients for Logistic Regression

	Amount test	
N = 268	Above Average	Below Average
Above Average	51	118
Below Average	91	8

Figure 3.2 Confusion Matrix for Logistic Regression

```
> cv.model
$size
[1] 4 3 2 1

$dev
[1] 100 100 437 437

$k
[1] -Inf    0  209  218

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Figure 4.1 Cv. model

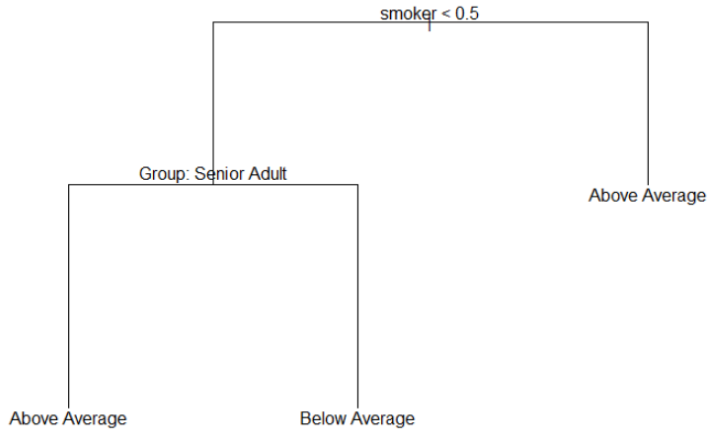


Figure 4.2 Pruned tree

	Amount test	
Prunetree.pred	Above Average	Below Average
Above Average	129	8
Below Average	13	118

Figure 4.3 Confusion Matrix for Classification Tree

	Amount test	
Yhat.rf	Above Average	Below Average
Above Average	129	8
Below Average	13	118

Figure 5.1 Confusion Matrix for Random Forest

importance(rf.in)	MeanDecreaseAccuracy	MeanDecreaseGini
sex	1.874534	5.086366
bmi	6.741533	53.106462
children	1.736063	16.294336
smoker	145.360231	156.065903
region	4.084009	12.821123
Group	121.600263	179.718045

Figure 5.2 Importance() plot for Random Forest

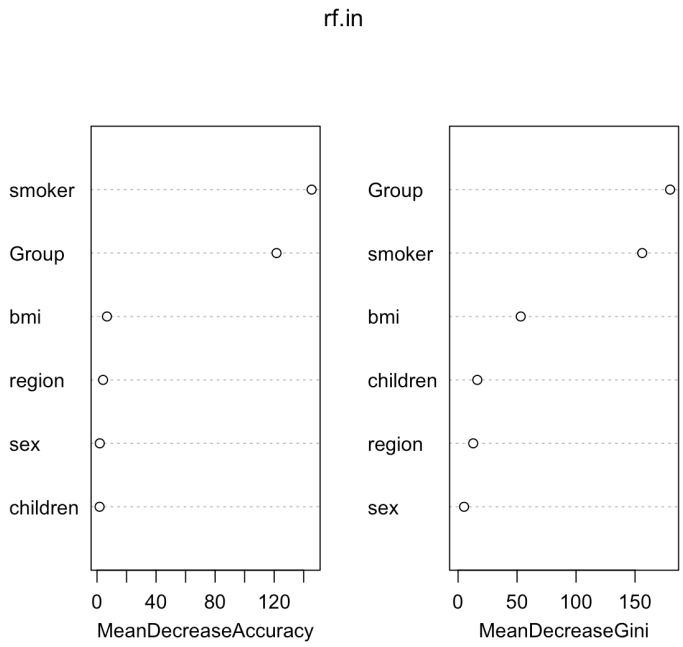


Figure 5.3 VarImpPlot() for Random Forest

References

"RandomForest." *Function | R Documentation*, www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance.

"Random Forests Leo Breiman and Adele Cutler." *Random Forests - Classification Description*, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.