Capstone Proposal

Julia Sheriff

8/16/2018

# Capstone Proposal

### Healthcare Charges of Varied Populations

##### By Julia Sheriff

#### **The Problem**

I would like to examine the ways in which various health variables impact health insurance charges in the United States. Variables include age, sex, BMI, number of children, smoking status, and region. I will explore the relationships of these variables with health insurance charges. Being able to estimate an individuial’s health insurance charges enables insurance providers to create plans that are both financially sound and ethical.

#### **The Client**

The clinet would be private insurance companies or the government. Both insurance companies and the government require cost analysis in order to develop insurance plans that are financially sound, maximizing profits while providing ethical reimbursement. Understanding which variables have the strongest impact on healthcare will assist in establishing ideal premium prices and service reimbursement rates.

#### **The Data**

The dataset is available at <https://www.kaggle.com/mirichoi0218/insurance/home>.

Health Variables:

|  |  |
| --- | --- |
| Variable | Description |
| Age | individual’s age in years |
| Sex | insurance contractor gender: female, male |
| BMI | Body mass index: metric that shows weight relative to height, (weight in kg / heght in m^2). Ideally 18.5 to 24.9. |
| Children | Number of children covered by health insurance, Number of dependents |
| Smoker | Smoker or Non-smoker |
| Region | Beneficiary’s US residental area: northeast, southeast, northwest, southwest |
| Charges | Individual medical costs billed by health insurance |

data4 <- read.csv("capstone\_data.csv", header=TRUE)  
  
head(data4)

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622

summary(data4)

## age sex bmi children smoker   
## Min. :18.00 female:662 Min. :15.96 Min. :0.000 no :1064   
## 1st Qu.:27.00 male :676 1st Qu.:26.30 1st Qu.:0.000 yes: 274   
## Median :39.00 Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.66 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000   
## Max. :64.00 Max. :53.13 Max. :5.000   
## region charges   
## northeast:324 Min. : 1122   
## northwest:325 1st Qu.: 4740   
## southeast:364 Median : 9382   
## southwest:325 Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

#### **The Approach**

My approach encompasses the following:

1. Data wrangling and cleaning
   * Account for any missing values or outliers
2. Exploratory Data Analysis
   * Use a combination of inferential statistics and data visualization to identify trends between health charges and health variables
     + Determine significant variables
     + Identify trends and correlations between variables
3. Machine Learning
   * Further examine the correlation between health charges and health variables using classification techniques and/or regression models
4. Data visualization and report out
   * Compile findings into deliverables (listed below)

#### **Deliverables**

The deliverables will consist of a report on my findings, a slide deck, and the corresponding R code used for data analysis. The deliverables will be submitted and published on GitHub.