Assignment: ASSIGNMENT 8.3 Final Project Step 1

Name: Shekhar , Manish

Date: 2021-05-12

1. Introduction

Reasearch Topic: Medical insurance costs

Health insurance provides important financial protection in case you have a serious accident or sickness.People without health coverage are exposed to these costs. This can sometimes lead people without coverage into deep debt or even into bankruptcy.

It's easy to underestimate how much medical care can cost:

1. Fixing a broken leg can cost up to $7,500

2. The average cost of a 3-day hospital stay is around $30,000

3. Comprehensive cancer care can cost hundreds of thousands of dollars

Having health coverage can help protect you from high, unexpected costs like these.

Thus, it is important that everyone, all the time, have affordable health insurance

regardless of where they work, their income, their age, or their health status

Affordable health insurance is the key to a productive work force, small business

innovation, and the economic as well as health security of our nation's families.

It is important to reaserch and understand components of rising health care costs

and propose probable changes to keep health insuance affordable for all.

Research can help understand medical cost -

1. variation by various age groups

2. variation amongst gender

3. variation by BMI (body mass index)

4. variation by number of dependents

5. variation by smoking habits

6. variation by US region etc.

Reasearch can help -

1. National, state, and local healthcare bodies to draft appropriate healthcare

policies to keep the medical insurance affordable to larger population with variety

of conditions.

2. Insurance companies to market policies and provide customizations to consumers

based on certain demographic behaviors.

3. Consumers to understand regional medical costs and affordability and consumer

behavior.

It's data science problem because it involves -

1. Data collection

2. Data cleansing

3. Data transformation

4. Data visualization

5. EDA - Exploratory data analysis

6. Modeling and Prediction

7. Validation and generalization

8. Integration and implementation

2. Research Questions

1. What is the average cost of health insurance in US by diffrent age groups?

2. What is the average cost of health insurance in US by gender?

3. What is the effect on health insurance cost by variation in BMI?

4. What is the effect on health insurance cost by number of dependents?

5. Why is the average health insurance cost varies in different US regions?

6. Predict the health insurance cost for given gender, age, BMI,

number of dependents, smoking habits, region etc.

7. What is the effect on health insurance cost by change in smoking habits?

8. What US region has most obese cases?

9. What US region has most smokers?

10. How is distribution of age groups amongst different US regions?

11. How is distribution of gender amongst different age groups and in different US regions?

12. Is there any correlation between BMI and smokers?

13. Do we see spike in any region in number of dependents?

14. What other factors can effect the insurance cost?

15. How much on an average American's pay for health insurance?

16. Which US region have most people with BMI less than average?

3. Approach

Approach involves analyzing data to discover correlations, patterns and create

machine learning model to predict cost of health insurance based of various

factors i.e. age, gender, bmi, region, smoking habit, number of dependents etc.

4. How does approach addresses the problem fully or partially?

Approach targets to give enough inputs to be able to address the problem completely.

It will help uncover various data patterns to answer multiple research questions.

It will help understand cause and effect relationship between health insurance cost

and various other factors i.e. age, gender, bmi, region, smoking habit, number of dependents etc.

It also intends to develop a model to predict health insurance cost given

various variables.

5. Data

```r
# Insurance data files for United States, one each for each region
ins_data_southwest <- read.csv("insurance_southwest.csv")
ins_data_southeast <- read.csv("insurance_southeast.csv")
ins_data_northwest <- read.csv("insurance_northwest.csv")
ins_data_northeast <- read.csv("insurance_northeast.csv")

# Combining the insurance data files into one data frame
# I manually inspected and they are all in same structure and thus can be combined using
# rbind into one data frame
ins_data <- rbind(ins_data_southwest,
                  ins_data_southeast,
                  ins_data_northwest,
                  ins_data_northeast)

# checking structure of the data
str(ins_data)
```

```
## 'data.frame':    1338 obs. of  8 variables:
##  $ X       : int  1 13 16 19 20 22 30 31 33 35 ...
##  $ age     : int  19 23 19 56 30 30 31 22 19 28 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 34.4 24.6 40.3 35.3 32.4 36.3 35.6 28.6 36.4 ...
##  $ children: int  0 0 1 0 0 1 2 0 5 1 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southwest" "southwest" "southwest" ...
##  $ charges : num  16885 1827 1837 10602 36837 ...
```

```r
# These datasets are inspired by the book Machine Learning with R by Brett Lantz. The data contains med

# Column definition
# -----------------
# age: age of primary beneficiary
# sex: insurance contractor gender, female, male
# bmi: Body mass index, providing an understanding of body, weights that are relatively high or low rel
# children: Number of children covered by health insurance / Number of dependents
# smoker: Smoking
# region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
# charges: Individual medical costs billed by health insurance

# Download link - https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv

# Information about how missing data values are recorded or how they were imputed is not provided.
# Checking for missing values in the data set
apply(ins_data, 2, function(x) any(is.na(x) | is.infinite(x)))
```

```
##        X      age      sex      bmi children   smoker   region  charges 
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
# There no missing values in the data set.
```

6. Packages needed for the project

Packages for data transformation

1. dplyr

2. purrr

Packages to Regression diagnostics

1. QuantPsyc - To get standard regression coefficients

2. car - Use durbinWatsonTest() to test the assumption of independent error

3. lmtest - Use dwtest() to test the assumption of independent error

Package for interactive plotting, model fitting, and stats about data

Rcmdr

Packages for data visualization and visual evaluation

1. ggplot2 - Useful to plot various charts to evaluate assumptions of linear regression

2. qqplotr - Useful to plot various charts to evaluate assumptions of linear regression

7. Questions for future steps

1. We delve deep into linear regression this week and definitely touched a variety of topics including linear regression diagnostics, fitting a linear model, selecting parameters, generlizing the model etc and associated statistical measures. One thing is definitely needed is more practice. Taking up different datasets and getting dirty while applying these concepts.

2. We learned about linear model assumptions and how to measure them. But we did not clearly cover what to do when each of these assumptions fail i.e. what are our options. Further reading and exploration on this topic is needed.

3. Looking forward to learn logistic regression as well. I am not sure as of now if I will see a use case to in this current topic I have chosen to be able to apply logistic regression. May be along additional consumer behavioral features we can use logistic regression classification to predict whether a lead will get converted or not i.e. will someone buy an insurance or any product or not.