machine learning approach

Julia Sheriff

10/28/2018

GENERAL IDEA OF DATASET

```
health_charges_clean <- read.csv("health_charges_clean.csv")
str(health_charges_clean)
```

```
## 'data.frame':    1338 obs. of  9 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age       : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex       : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi       : num  27.9 33.8 33 22.7 28.9 ...
##  $ bmi_factor: Factor w/ 4 levels "healthy_weight",..: 3 2 2 1 3 3 2 3 3 3 ...
##  $ children  : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker    : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region    : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges   : num  16885 1726 4449 21984 3867 ...
```

```
head(health_charges_clean)
```

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

PREPPING MY VARIABLES

- CHARGES - dependent variable for all predictions
  - continuous
  - CREATE 2 binary vars (above .75 IQR, below .75 IQR)
- Age
  - continuous
  - CREATE 5 binary vars (10 year groups (18-25 counted as one bracket))
- Bmi
  - continuous
  - categorical (I now have 4 categories, but am adding two more (breaking up obese into 3 categories according to CDC)).
  - CREATE 6 binary vars
- Sex
  - binary
- Children
  - categorical
  - CREATE 5 binary vars
- Smoker
  - binary
- Region

  - categorical
  - CRETE 4 binary vars

LINEAR REGRESSIONS:

- INITIAL MODELS
  - BMI as continuous
  - QUESTION: According to mean graph, linear relationship changes when BMI is under 30, 30-35, and over 35. How do I do an accurate linear regression? Should I subset the data and do three separate linear regressions for BMI under 30, 30-35, and over 40?
  - Age as continuous
  - Age and BMI combined
  - QUESTION: Again, should I subset the data according to the changes in the linearity of BMI?
- ASSESSMENT:
  - See which regression shows the highest R^2

LOGISTIC REGRESSIONS:

- INITIAL MODEL includes all independent variables:
  - Smoker
  - Bmi (categorical)
  - Sex
  - Region
  - Children
  - Age
- ASSESSMENT:
  - See which are strong predictors
  - Create new and improved model
  - Calculate probabilities for high charges from each facet used in final model

CLUSTERING:

- VARIABLES- all variables as binary (including charges).
- PROCESS 1:
  - Heigharachial dendrogram if R allows, to choose # of clusters
  - Apply clustering
- PRCESS 1 ASSESSMENT: Find percentage of "high" charges found in each group
- (PROCESS 2):
  - Learn how to do something meaningful using the last kmeans exercise I couldn't do...