machine learning approach

Julia Sheriff

10/28/2018

PREPPING MY VARIABLES

- CHARGES - dependent variable for all predictions
  - continuous
  - binary (above .75 IQR, below .75 IQR)
- Age
  - continuous
  - binary (5 facets in 10 year groups (18-25 counted as one bracket))
- Bmi
  - continuous
  - binary (6 facets of BMI)
- Sex
  - binary (2 facets)
- Children
  - binary (6 facets)
- Smoker
  - binary (2 facets)
- Region
  - binary (4 facets)

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 35 and BMI over 35?
  - Age as continuous
  - Age and BMI combined
  - Again, should I subset the data according to the changes in the linearity of BMI?

- ASSESSMENT:
  - See which regression shows the highest R^2 and precision

LOGISTIC REGRESSIONS:

- INITIAL MODELS:
  - 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
  - Find percentage of "high" charges found in each group
- (PROCESS 2):
  - Learn how to do with with kmeans)