FEATURE SELECTION & FEATURE CREATION

LOADING THE CENSUS DATASET

- LOAD DATA FILE

WHAT QUESTION ARE WE ASWERING?
SUPERVISED OR USUPERVISED?

IF SUPERVISED, WHAT IS OUR TARGET VARIABLE?

FEATURE ENGINEERING

LAST CLASS WE SAW THE NECESSARY VARIABLE TRANSFORMATIONS:

- NA TREATMENT
- OUTLIER REMOVAL
- CATEGORICAL ENCODING
- TEST/TRAIN SAMPLING

-ADDING VARIABLES (FEATURE CREATION)
- VARIABLE SELECTION (FEATURE SELECTION)

FEATURE CREATION

- NUMERICAL TRANSFORMATION;
- CATEGORICAL TRANSFORMATION;
- TARGET VARIABLE TRANSFORMATION;

VARIABLE COMBINATION

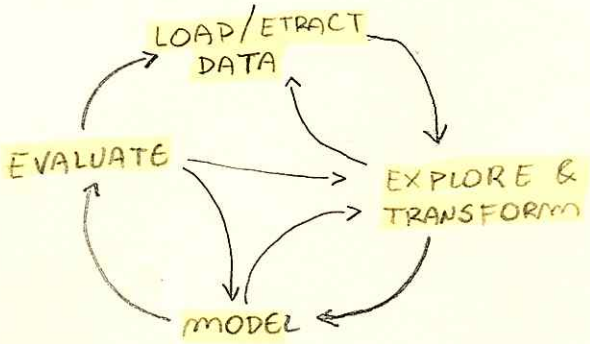
- INTERACTION BETWEEN CATEGORICAL & CONTINOUS VARIABLES;
- INTERACTION BETWEEN CONTINOUS VARIABLES;
- INTERACTION BETWEEN CATEGORICAL VARIABLES.

PCA AS PRE-PROCESSING

FEATURE SELECTION: REDUCING COLINEARITY

- MANUAL SELECTION: DOMAIN KNOWLEDGE, PROBLEM REQUIREMENTS
  * GREATER IMPORTANCE ON EDA!
- PCA & NMF;
- LASSO REGRESSION / RIDGE REGRESSION;

DATA ANALYSIS IS AN ITERATIVE PROCESS:



COMMON NUMERICAL TRANSFORMATION

$x > 0$:

$$T(x) = \log(x)$$

$$T(x) = \sqrt{x}$$

$0 < x < 1$:

$$T(x) = \frac{\log(x)}{1 - \log(x)}$$

$-\infty < x < \infty$:

PowerTransformer()
QuantileTransformer()
qcut()
cut()

CATEGORICAL TRANSFORMATION

- CATEGORICAL GROUPING
- CATEGORY CREATION

COMMON TARGET VARIABLE TRANSFORMATION

$y > 0 \rightarrow \log(y)$
PowerTransformer

$0 < y < 1 \rightarrow logit(y) = \frac{\log(y)}{1 - \log(y)}$

$a < y < b \rightarrow \bar{y} = \frac{y - a}{b - a}$

$$\frac{\log(\bar{y})}{1 - \log(\bar{y})}$$