Predicting the need for medical imaging on patients' information

Group 2: Keyu Chen, Yinuo Chen, Rongqian Zhang

Dataset

- Ambulatory Health Care Data
- The National Hospital Ambulatory Medical Care Survey (NHAMCS)
- **Patient** characteristics : age, sex, race, and ethnicity, and so on
- **Visiting** characteristics : temperature, heart rate, systolic blood pressure, pulse oximetry, and so on
- **Text** characteristics : reasons for visit and possible causes of disease
- Response variable : **ANYIMAGE** (X-ray and CT scan)

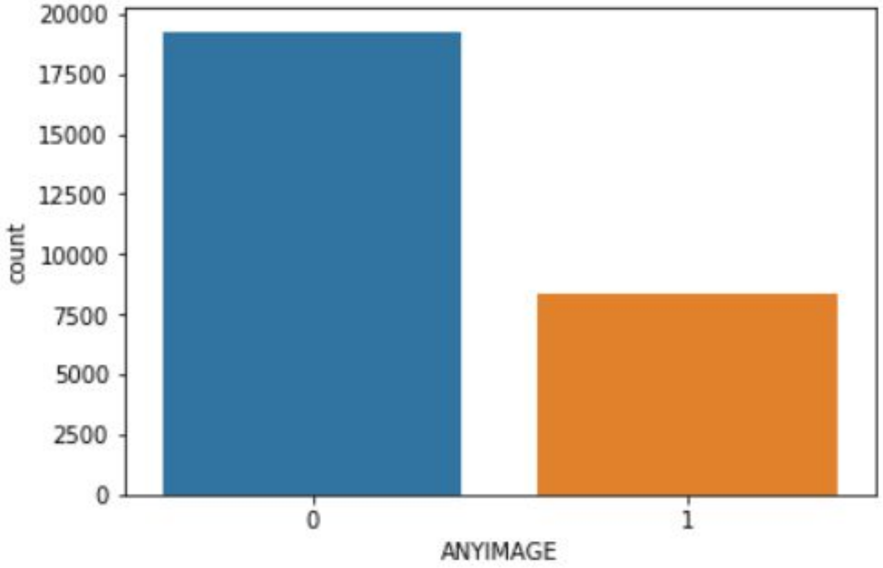
Source: https://www.cdc.gov/nchs/ahcd/index.htm

Question

Build a model to predict whether patients need any medical imaging

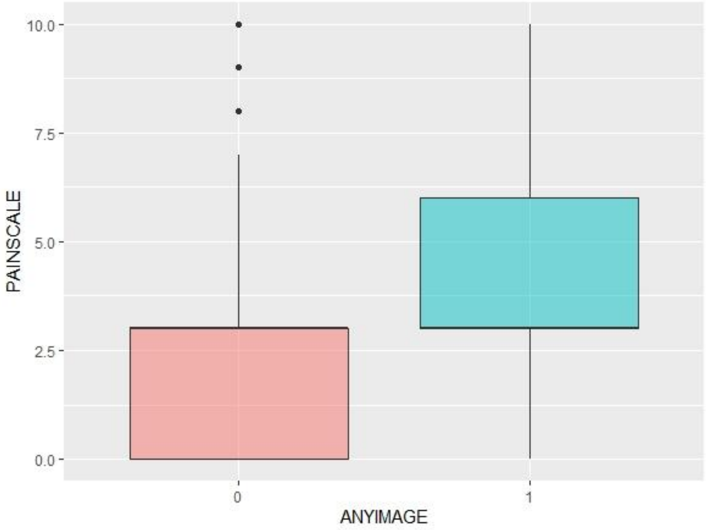
EDA - Response

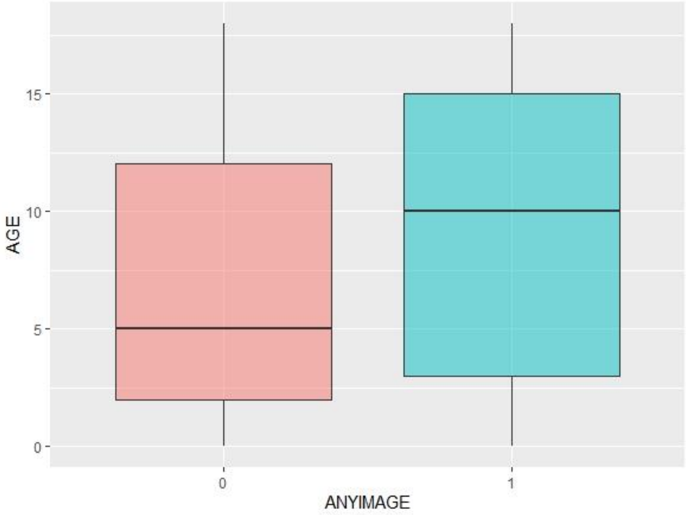
- 2012 - 2016: 27,665 observations
- Positive : Negative = 1 : 2.3



EDA - Structured Data

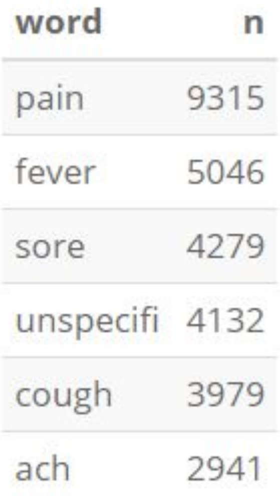
- Some variables have significantly different distribution in two groups
- Ex: Pain scale, age



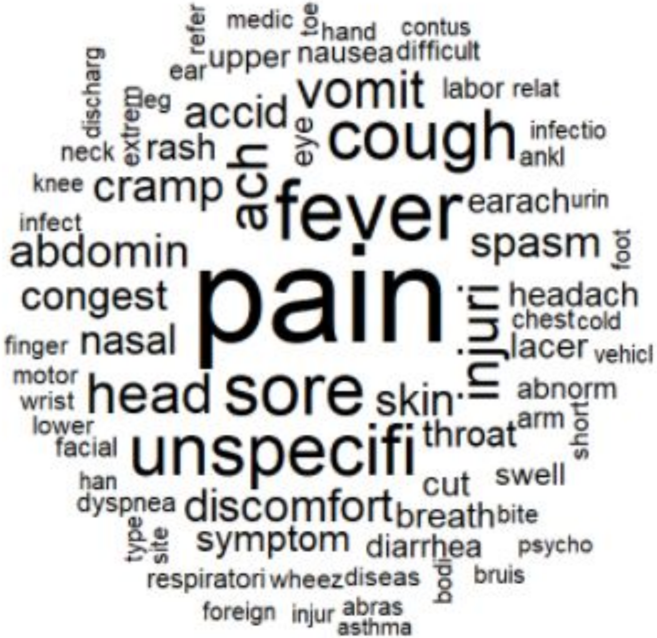


EDA - Text Data

- Data cleaning
  - Remove urls, punctuations, single letter
  - Remove stop-words
  - Lowercase
- Reasons for visiting
  - Most common words
  - Word cloud

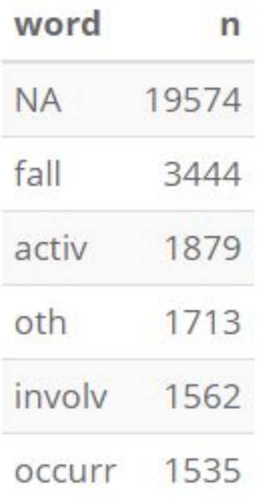


word	n
pain	9315
fever	5046
sore	4279
unspecifi	4132
cough	3979
ach	2941

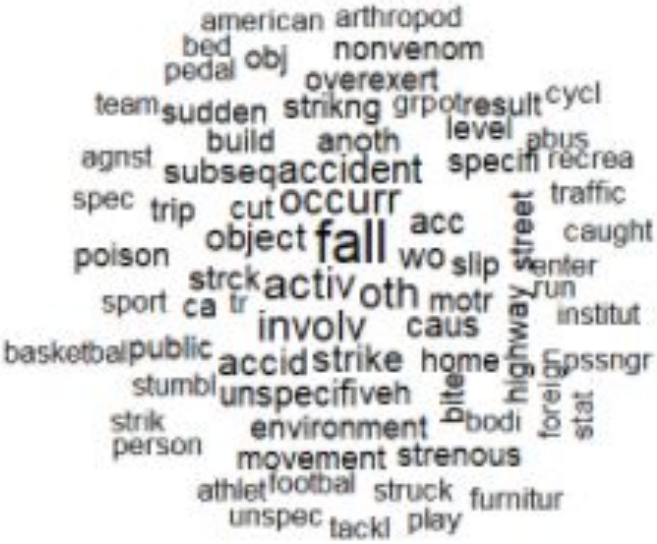


EDA - Text Data

- Possible causes of disease
  - Most common words
  - Word cloud

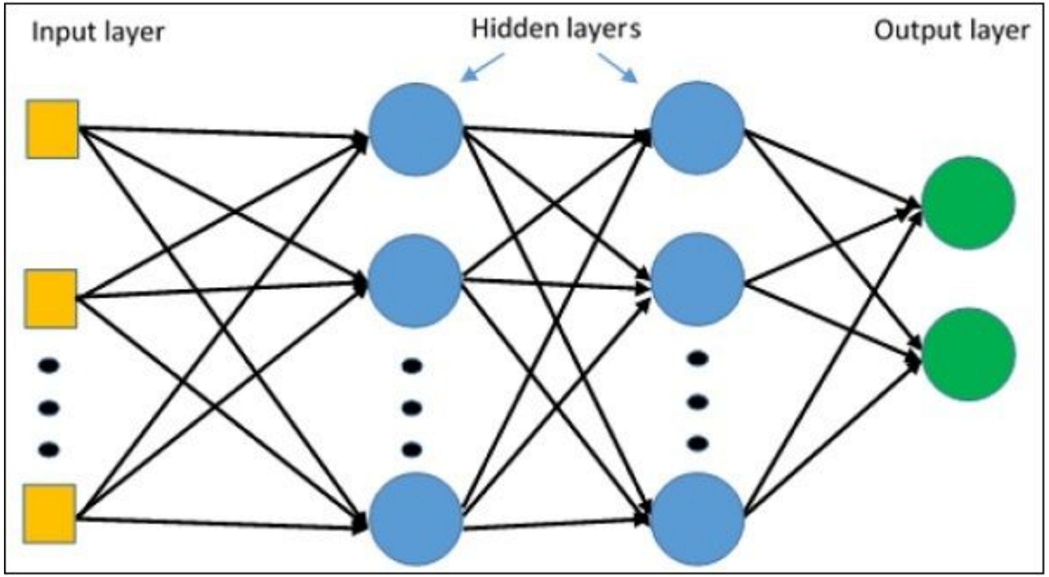


word	n
NA	19574
fall	3444
activ	1879
oth	1713
involv	1562
occurr	1535



Model - structured variables

- Multilayer perceptron



Source: TensorFlow - Multi-Layer Perceptron Learning

Model - text variable

- Word embedding - GloVe
  - Global Vectors for Word Representation
  - Based on word-word co-occurrence matrix
  - Similar words will have similar representation

The cat sat on the mat.
The dog sat on the mat.

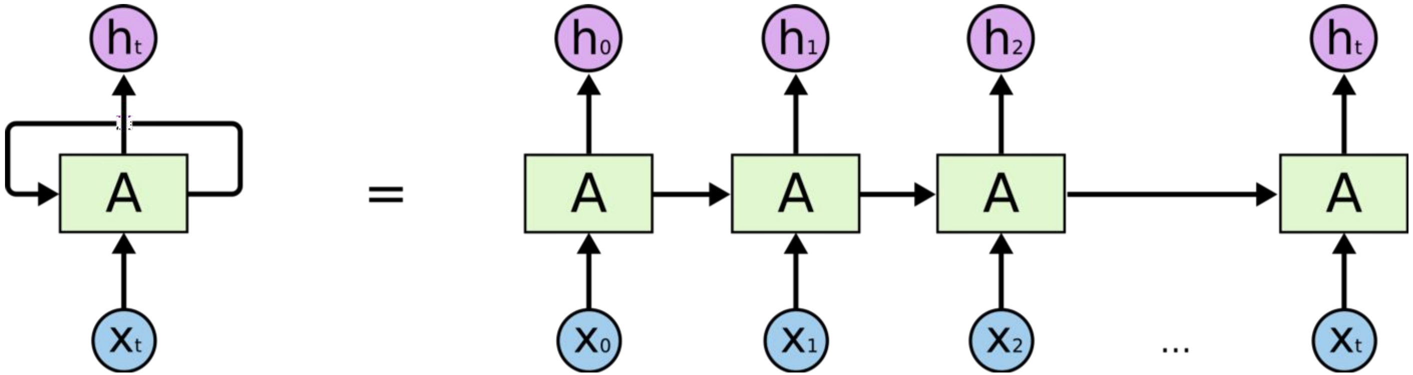
co-occurrence matrix
with a window size of 1.

	the	cat	sat	on	mat	dog
the	0	1	0	2	2	1
cat	1	0	1	0	0	0
sat	0	1	0	2	0	1
on	2	0	2	0	0	0
mat	2	0	0	0	0	0
dog	1	0	1	0	0	0

Source: https://nlp.stanford.edu/projects/glove/

Model - text variable

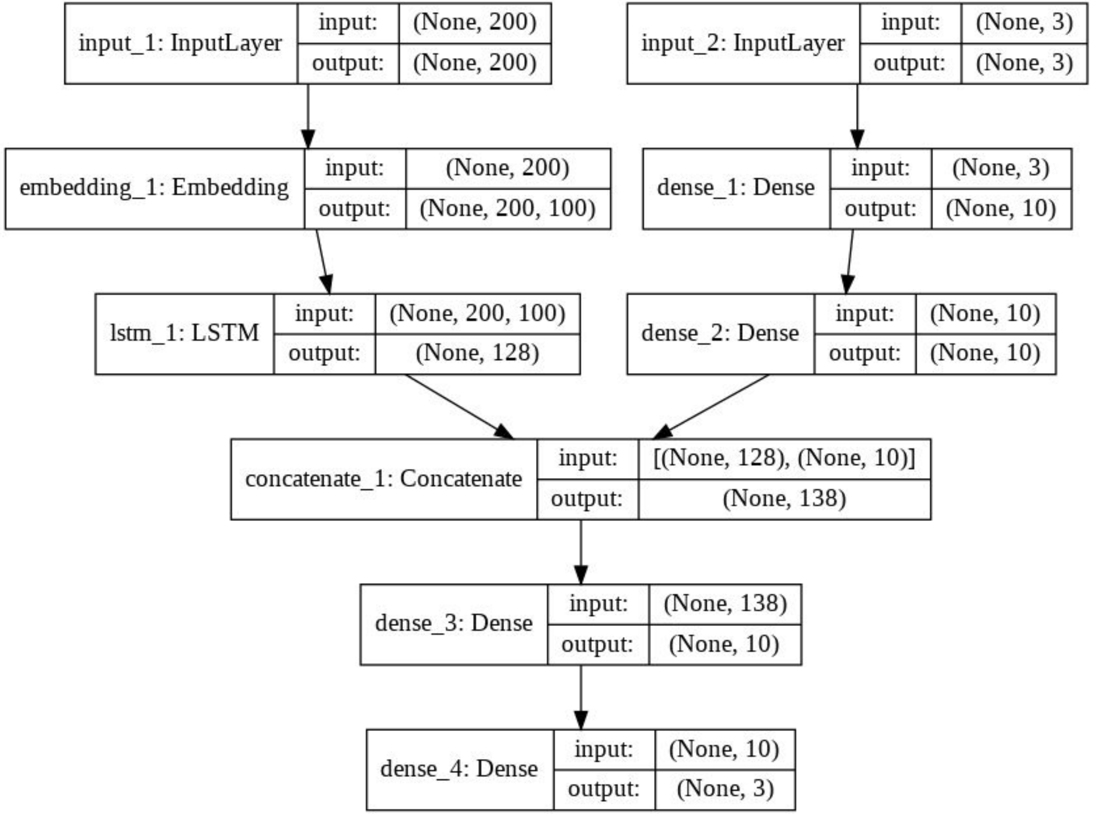
- LSTM (Long short-term memory)
  - Similar to the way human read sentence
  - LSTM read words sequentially
  - At each time stamp, it uses current word and the activation from last timestamp as input



Source: https://medium.com/datathings/the-magic-of-lstm-neural-networks-6775e8b540cd

Model - mixed variables

- Hybrid neural network
    - Concatenating MLP and LSTM



input_1: InputLayer | input: (None, 200) | output: (None, 200)

input_2: InputLayer | input: (None, 3) | output: (None, 3)

embedding_1: Embedding | input: (None, 200) | output: (None, 200, 100)

dense_1: Dense | input: (None, 3) | output: (None, 10)

lstm_1: LSTM | input: (None, 200, 100) | output: (None, 128)

dense_2: Dense | input: (None, 10) | output: (None, 10)

concatenate_1: Concatenate | input: [(None, 128), (None, 10)] | output: (None, 138)

dense_3: Dense | input: (None, 138) | output: (None, 10)

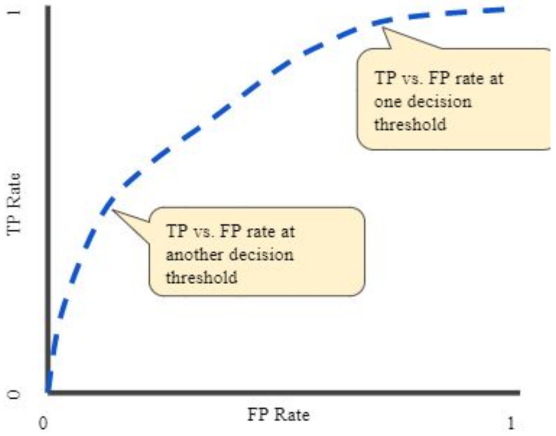
dense_4: Dense | input: (None, 10) | output: (None, 3)

Results

- Evaluation
- AUC (Area Under The Curve)

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$



Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Thank you!