Determining the Accuracy of Probabilistic Scoring

Client:

Alan Malik

Patient Tools
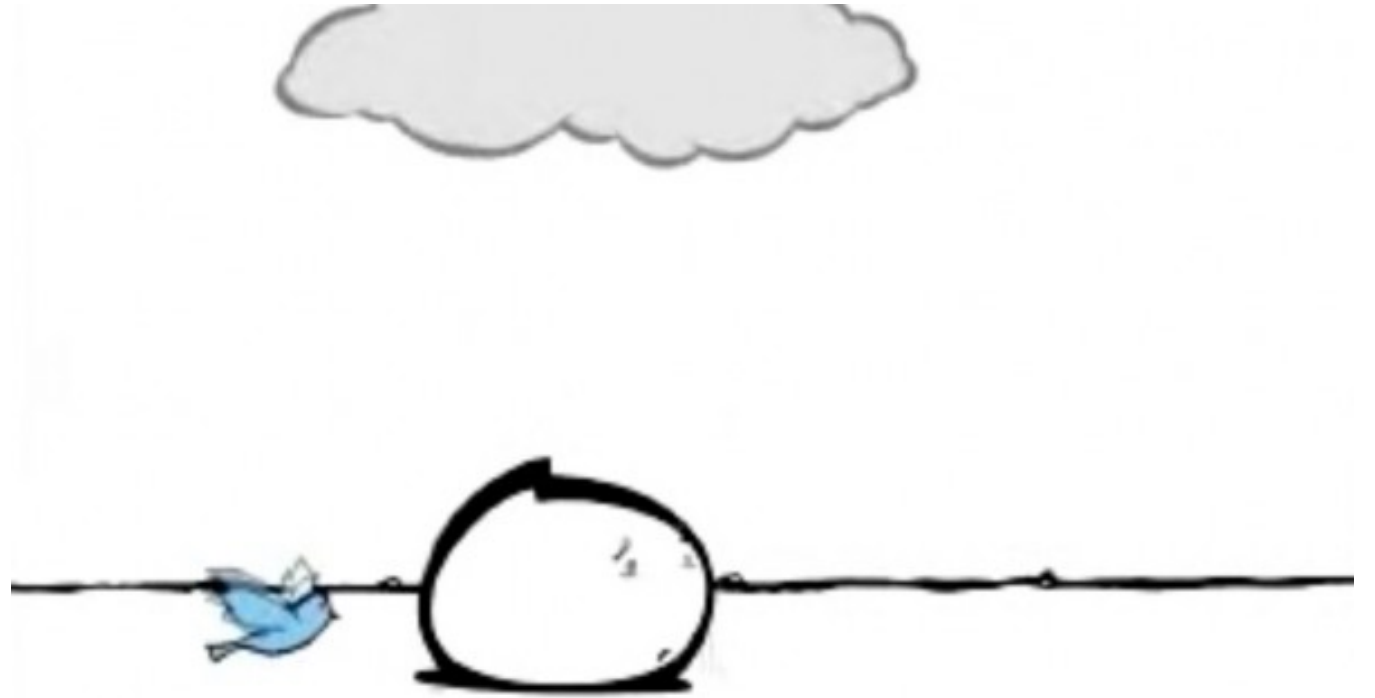
Lee Panter

ARE YOU SAD?

Do your Patient Health Questionnaire (PHQ) summaries give you feelings of:

- Inaccuracy

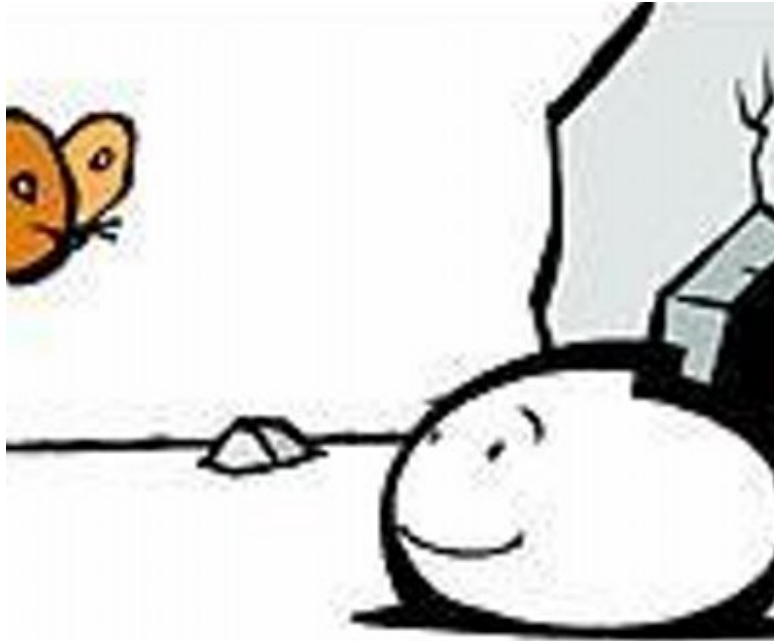- Low specificity

- Low sensitivity

These may be symptoms of a more serious condition:

Insufficient Statistic (itus)™

Talk to your statistician today about:

PROBABILISTIC SCORING

* Results may vary, Use as directed, May cause drowsiness-do not operate machinery until you know how this treatment affects you...

Client Proposal-Problem Statement

- Questionnaires are standard approach to preventative screening for mental health disorders

- Designed to be:
  - Easy to use, track and record
  - Quick (maximum average of 6.5 minutes)
  - Reduced to minimum necessary questions, prioritizing answers with higher risk for further evaluation

- Standard Testing Protocol
  - PHQ-2 → PHQ-9 → QDP → Mental Healthcare Specialist
  - PHQ-2 (Patient Health Questionnaire)
    - First two questions of PHQ-9
    - Scores are 0-6
    - Scores higher than 2 considered "at risk"

- PHQ-9
  - Scores are 0-27
  - Scores higher than 9 considered "at risk"
- QDP (Quick PsychoDiagnostic Panel)
  - Dynamic decision-tree protocol based on DSM-V
  - Can classify 11 disorders

What's the Problem?

- Poor quality statistics used for classification
- Inferior performance due to poor classification caused clinical fatigue
- One study found that after using the PHQ-2, providers did not refer a positive result 95% of the time!

Client Proposal-Solution

Collected Data:
- 2,495 total cases (observations)
- PHQ-9 & QDP results
- Collected over six months
- Federally Qualified Health Center in Montana
- QDP administration was randomized
  - Before/after taking PHQ-9
  - Some cases only took PHQ-9

Definitions
- E = percent of cases endorsing a specific item answer
  - (e.g percent of cases that answered 0 on q1)
- C = outcome class
  - (e.g. the outcome class of case 1 is "Major Depression")

Probabilistic Scores
- Iterative process
  - Updated after another question is answered
- A single score
  - Probability that the given series of answers dictates a certain classification
  - PHQ-9 has three categories to classify therefore each score is three numbers

The Process
1. Use given data to calculate the response probabilities: $p_{ijk} = \frac{P(E_{ijk}|C_i)*P(C_i)}{P(E_{ijk})}$

Where
- i = { Not Clinically Depressed, Sub-threshold Depression, Major Depression}
- $j = 1,2,\ldots,9$ and
- $k = 0,1,2,3$

2. Upon observing answer to next question, update distributions for E and C, and re-calculated p_{ijk}
3. Repeat (1) ← → (2) until a single component probability of p_{ijk} passes acceptable threshold values

Client and Consultant Goals

Main Consultant Goals

- Mathematically prove that Probabilistic scoring is more accurate than conventional scoring
  - "…show that information is not lost when applying Bayes' Theorem"
  - Establish minimum sample size needed for Probabilistic Scoring error to be smaller than that of conventional scoring.

- Mathematically prove that probabilistic scoring derived from a conventional scored validation dataset is essentially as accurate as using the original validation dataset and therefore still more accurate than conventional scoring

Initial Steps

- Literature Review
  - Item Response Theory
  - Bayesian sufficiency
- Discuss areas of confusion with client and class
- Obtain client coding methods including data used to generate results
- Possibly inquire into previous proof attempts made, and other resources that might be useful

Client Goals

- Patent Concept (Complete)
- Educate medical practitioners (spreading the word)
- Getting the concept accepted as Clinically appropriate
- Obtain large PHQ-9 data set to test error convergence based upon error size

I have a question…

- There are several times in the paper where accuracy, false positives, false negatives,…, etc. are referenced. How are you calculating these values?
- What are the "cut scores" for each test you implemented? How are/were they determined?
- Is there a standard decision-tree you have been using that minimizes the error? Have you tried re-ordering the responses?
- Along a similar line, what sort of case/question independence assumptions are you making?
  - Are the cases independent?
  - How about the questions (assuming no decision tree)?
- Are you assuming that your sample is representative of your population of interest?
  - Will your database of weights be continually reset (i.e. go back to the original settings after each test), or will each test also be factored in?
- Are asymptotic results and/or asymptotic-dependent results acceptable?
- What parametric assumptions were made in your analysis? Is a proof that is specific to certain parametric values desirable?

You should stop Probabilistic Scoring if:

- You are allergic to math
- You dislike being correct

Seek immediate Statistical Assistance if you experience a positive predictive value greater than one

- For external use only
- Rinse mouth thoroughly after each use
- May cause drowsiness
- Package is NOT child resistant

THANK YOU!!