

# Final Project Outline

*Lee Panter*

---

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background & Context necessary for project . . . . .	2
1.2	Data . . . . .	2
1.2.1	Data Origin: . . . . .	2
1.2.2	Data Properties: . . . . .	2
1.3	Classification Methods . . . . .	3
1.3.1	Traditional Classification . . . . .	3
1.3.2	Probabilistic Classification . . . . .	3
1.4	Consultation goals and deliverables . . . . .	4
1.4.1	Client's Specified Goals: . . . . .	4
1.4.2	Analysis Goals: . . . . .	4
1.4.3	Deliverables: . . . . .	5
<b>2</b>	<b>Model and Methods</b>	<b>5</b>
2.1	Quantifying Accuracy . . . . .	5
2.2	Cross Validation Processing: . . . . .	5
2.3	Optimal Convergence Criterion . . . . .	8
2.4	Simulated Validation Outcomes . . . . .	8
2.4.1	Inferential Simulations . . . . .	8
2.4.2	Probabilistic Algorithm Simmulations . . . . .	9
2.4.3	Unsupervised Learning Classifications . . . . .	11
2.4.3.1	Kmeans and Heirarchical Clustering . . . . .	12
2.4.3.2	Self Organizing Maps . . . . .	13
<b>3</b>	<b>Analysis and Results</b>	<b>14</b>
3.1	Inferential Outcome Simulation . . . . .	14
3.2	Probabilistic Simulations . . . . .	15
<b>4</b>	<b>Discussion</b>	<b>16</b>
<b>5</b>	<b>Appendix</b>	<b>16</b>
<b>6</b>	<b>References</b>	<b>16</b>

---

---

# 1 Introduction

## 1.1 Background & Context necessary for project

- The Patient Health Questionnaire-Nine (PHQ9)
  - Self-administered module
  - Can be administered by medical staff
  - Nine questions, answers correspond to the frequency with which respondents feel certain feelings
  - Used for screening, monitoring, and grading severity of depressive symptoms related to nine criteria outlined by the Diagnostic and Statistical Manual of Mental Health Disorders (DSM-IV) [1]
  - Developed by Dr. Robert J. Spitzer, Dr. Janet B.W. Williams, Dr. Kurt Kroenke in 1999 with a grant from Pfizer [2]
  - Response sets are classified into a discrete set of categories corresponding to depression symptom magnitude
  - Three categories corresponding to: “Not clinically Depressed”, “Sub-threshold Depression”, and “Major Depression” will be used for this analysis

## 1.2 Data

### 1.2.1 Data Origin:

- Federally Qualified Health Research Center in Montana [3]
- Collected over six month timeframe
- Tablet-administered PHQ9s and Quick Diagnostic Panels (QDPs)

### 1.2.2 Data Properties:

- Technically a Randomized Control Design based upon assignment of QDPs, but observational based upon PHQ9 data
- 2495 observations
- 286 variables: PHQ9 data, QDP variables, Control variables, Demographic variables, Record-keeping (time, date, etc) variables
- De-identified, and left without validation outcome measurement
- No theoretically “correct” classification
- Each observation contains integer-value responses (0-3) for the nine PHQ9 questions
- No missing data
- Original test was administered sequentially, with all questions being asked
- Some respondents had taken the QDP first

## 1.3 Classification Methods

### 1.3.1 Traditional Classification

- Traditional classification class representations:
  - $\mathbf{C}^{TR} = (C_1^{TR}, C_2^{TR}, C_3^{TR})$
- Assign numerical quantities to each response (0-3) for questions (1-9).
  - An answer set corresponding to observation  $i = 1, \dots, N$  is represented as
  - $\mathbf{A}_i = \{a_q\}_{q=1}^9$  where  $i = 1, \dots, N$  and  $a_q \in \{0, 1, 2, 3\}$  for  $q = 1, \dots, 9$
- Calculate sum of assigned numerical quantities (0-27)
  - The sum of an answer set provided in observation  $i = 1, \dots, N$  is represented by:
  - $S_i = \sum_{q=1}^9 a_q$
- Use sum to classify observation
- Depression classes are distinguished by “threshold values”
  - let  $\alpha_1, \alpha_2 \in \{0, \dots, 27\}$  with  $\alpha_1 \leq \alpha_2$  be threshold values
  - We can define the Traditional class representations  $C_c^{TR}$  for  $c = 1, 2, 3$  as Level sets according to:
    - \*  $C_1^{TR} = \left\{ i \mid S_i < \alpha_1 \right\}$
    - \*  $C_2^{TR} = \left\{ i \mid \alpha_1 \leq S_i < \alpha_2 \right\}$
    - \*  $C_3^{TR} = \left\{ i \mid \alpha_2 \leq S_i \right\}$
- for this analysis, threshold values are:  $\alpha_1 = 7$  and  $\alpha_2 = 10$
- 88% accurate [4]
- 12% FP/FN causes issues with Clinical Fatigue, patient concern, and healthcare system burdens
- Outcomes provide limited information
- Need to take the entire PHQ9 in order to achieve results

### 1.3.2 Probabilistic Classification

- Possible benefits:
  - Reduced test-taking requirements (early convergence)
  - Increased relatability with other Mental Health disorders, and better result integration in general
  - Not dependent on numerical assignments: integer differences of outcomes less influential.
  - Training-sample accuracy increase possibility
- The algorithm:
  - Probabilistic classification class representation
    - \*  $\mathbf{C}^{PR} = (C_1^{PR}, C_2^{PR}, C_3^{PR})$
  - Let the question number be represented by:  $Q = q \in \{1, 2, \dots, 9\}$
  - Let the provided answer to a particular question be represented by:  $A = a \in \{0, 1, 2, 3\}$

- Given training set, a set of weights is calculated using the formula:

$$W_{AQ}^{CPR} = \frac{P(Q = q \mid C^{TR} = c)}{P(Q = q)} \quad (1.3.2 - 1)$$

- Suppose that  $\mathbf{A}_i = \{a_q\}_{q=1}^9$  represents a set of answers to the PHQ9, contained in observation  $i$ .
- For a provided confidence threshold value  $\gamma \in (0, 1)$ , the probabilistic scoring sequence is given recursively by:

$$\begin{aligned} (P_0^{(1)}, P_0^{(2)}, P_0^{(3)}) &= \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \\ (P_{q+1}^{(1)}, P_{q+1}^{(2)}, P_{q+1}^{(3)}) &= \frac{\left(P_q^{(1)}W_{a_q}^{(1)}, P_q^{(2)}W_{a_q}^{(2)}, P_q^{(3)}W_{a_q}^{(3)}\right)}{\sum_{j=1}^3 P_q^j W_{a_q q}^{(j)}} \end{aligned} \quad (1.3.2 - 2)$$

- If we define:

$$j_q^* = \max_j P_q^{(j)}$$

and

$$q^* = \min_q \{q : P_q^* > \gamma\}$$

then (provided that  $q^*$  exists), the probabilistic scoring classification is:

$$j_{q^*}^* = \max_j P_{q^*}^j$$

more specifically, it is:

$$C_{j_{q^*}^*}^{PR}$$

## 1.4 Consultation goals and deliverables

### 1.4.1 Client's Specified Goals:

“... mathematically prove that probabilistic scoring is more accurate than 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”

### 1.4.2 Analysis Goals:

- Compare Probabilistic Scoring accuracy to Conventional Scoring accuracy measured against simulated responses generated using information in PHQ9 data.
- Determine how accuracy comparisons vary as a function of training sample size.

### 1.4.3 Deliverables:

- Presentable results. Evidence of outcomes that can be shown to potential clients with the purpose of demonstrating the practical advantages of Probabilistic Scoring.
  - Relationship Diagrams. Visualizations that can be utilized to show informational gain when Probabilistic Scoring is implemented. Particular interest in relating outcomes to specific question answers.
- 

## 2 Model and Methods

### 2.1 Quantifying Accuracy

- Accuracy as a function of model value and validation outcome

$$\text{Accuracy} = f(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^N I(y_i = \hat{y}_i)$$

- where  $\mathbf{y} = (y_1, \dots, y_N)$  is the validation outcome
- $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_N)$  is the model value
- $I(y_i = \hat{y}_i) = \begin{cases} 0 & \text{if } y_i \neq \hat{y}_i \\ 1 & \text{if } y_i = \hat{y}_i \end{cases}$
- Accuracy as a function of model value and *simmulated* validation outcome

$$\tilde{\text{Accuracy}} = f(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^N I(\tilde{y}_i = \hat{y}_i)$$

- where  $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_N)$  is the simmulated validation outcome
- Simmulated accuracy calculation can be performed for Traditional and Probabilistic Classifications, and then compared:

$$\begin{aligned} \tilde{\text{Accuracy}}_{TR} &\sim \tilde{\text{Accuracy}}_{PR} \\ f(\tilde{\mathbf{y}}, \hat{\mathbf{y}}^{TR}) &\sim f(\tilde{\mathbf{y}}, \hat{\mathbf{y}}^{PR}) \end{aligned} \quad (2.1-1)$$

- This analysis will focus on estimating various values of  $\tilde{\text{accuracy}}$  using multiple methods of generating values of  $\tilde{\mathbf{y}}$

### 2.2 Cross Validation Processing:

- Estimating  $\tilde{\text{Accuracy}}$  using Cross Validation allows us to obtain an average-stabilized estimate of both Pscore and traditional classifications simmultanneously, and to establish the relationship between training sample size and accuracy

- For a fixed value of  $K \in \{2, \dots, N - 1\}$  we partition the full data set into  $K$  equal subsets, each of which has  $N_K$  observations sampled randomly without replacement from the original data, where:

$$N_K = \left\lfloor \frac{N}{K} \right\rfloor$$

- We now have  $K$  distinct data sets that have been subset from the original data, that we will combine into test-train pairs
- There are  $K$  different ways in which  $K - 1$  data sets can be chosen from  $K$  sets when order does not matter, for each of these  $K$  combinations we create a train-test combination by combining the  $K - 1$  selected sets to form the training set, and the remaining set to form the test set.
- We will denote these train/test pairs

$$(TR, TE)_j^k = (TR_j^k, TE_j^k)$$

for  $j = 1, \dots, K$

- From the definition of  $N_K$ , and the fact that each training set is a union of  $(K - 1)$  sets of length  $N_K$ , we can establish a relationship between training set length ( $|TR_j^k|$ -the number of observations in the  $j$ th training set) and the value of  $K$  from which the original test-train partitions were originally formed:

$$|TR_j^k| = (K - 1)N_K = (K - 1)\left\lfloor \frac{N}{K} \right\rfloor$$

- Similarly, we may define the length of the test set:

$$|TE_j^k| = N_K = \left\lfloor \frac{N}{K} \right\rfloor$$

- noting that:

$$\left\lfloor \frac{N}{K} \right\rfloor \leq \frac{N}{K}$$

We have

$$\begin{aligned} |TR_j^K| + |TE_j^K| &= (K - 1)\left\lfloor \frac{N}{K} \right\rfloor + \left\lfloor \frac{N}{K} \right\rfloor \\ &= K\left\lfloor \frac{N}{K} \right\rfloor \\ &\leq K\frac{N}{K} = N \end{aligned}$$

Implying that some observations may be left out of the CV process.

- We are interested in establishing a relationship between  $|TR_j^k|$  and *accuracy*, idealistically represented by:

$$\tilde{\text{accuracy}}(|TE_j^k|) = \tilde{\text{accuracy}}\left(\left\lfloor \frac{N}{K} \right\rfloor\right)$$

This formula establishes a connection between the value of  $K$  and *accuracy*. Demonstrating that by calculating the value of *accuracy* at varying the value of  $K$ , we may establish the relationship between *accuracy* and  $|TR_j^k|$  (i.e. Training Sample Length).

- We can calculate the calculate the classification accuracy for each train-test pairing within a specific value of K.

$$\tilde{accuracy} (TR, TE)_j^k = \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I(\tilde{y}_i = \hat{y}_i)$$

- and we will use the average of these values to represent the accuracy at a particular K-value

$$\begin{aligned} \tilde{accuracy} (TR, TE)^k &= \tilde{accuracy}_m (TR, TE)_\bullet^k \\ &= \frac{1}{K} \sum_{j=1}^K \left\{ \tilde{accuracy} (TR, TE)_j^k \right\} \\ &= \frac{1}{K} \sum_{j=1}^K \left\{ \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I(\tilde{y}_i = \hat{y}_i) \right\} \end{aligned}$$

- This process can be applied to Traditional and Probabilistic Scoring methods:

$$\begin{aligned} \tilde{accuracy}_m (TR, TE)^k &= \frac{1}{K} \sum_{j=1}^K \left\{ \tilde{accuracy}_m (TR, TE)_j^k \right\} \\ &= \frac{1}{K} \sum_{j=1}^K \left\{ \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I(\tilde{y}_i = \hat{y}_i^m) \right\} \end{aligned}$$

where  $m \in \{Pscore, Tscore\}$ ,  $\hat{y}^{Pscore}$  is a Probabilistic Scoring generated outcome, and  $\hat{y}^{Tscore}$  is a Traditional Scoring generate outcome

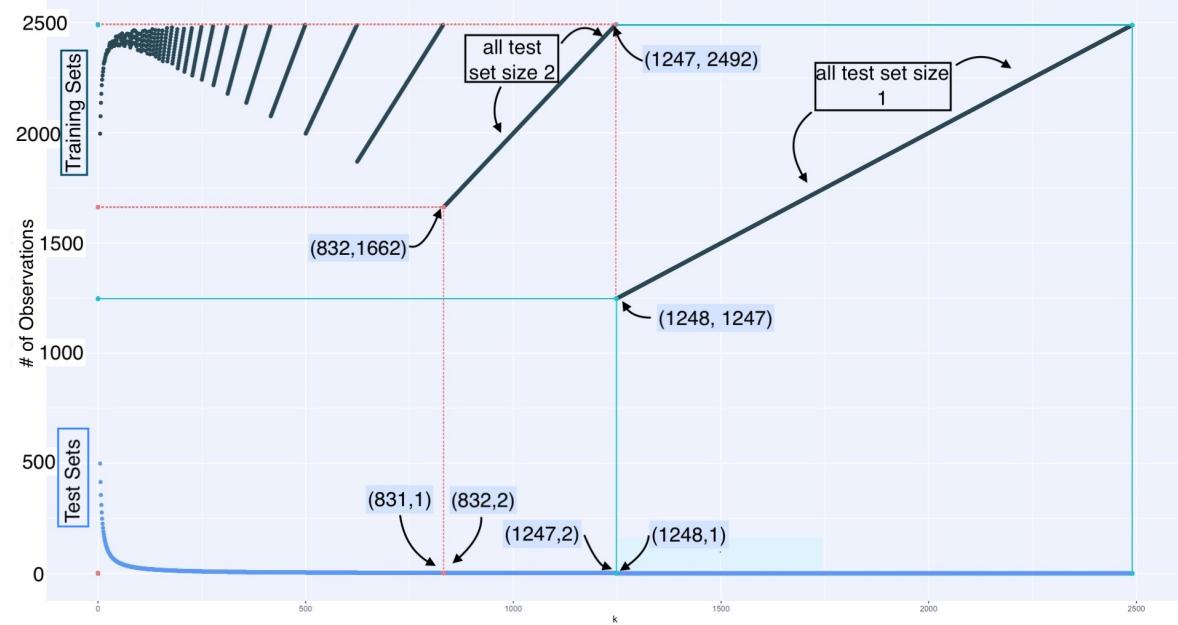
- It will also be applied to a multitude of different simulated validation outcomes

$$\begin{aligned} \tilde{accuracy}_m^h (TR, TE)^k &= \frac{1}{K} \sum_{j=1}^K \left\{ \tilde{accuracy}_m^h (TR, TE)_j^k \right\} \\ &= \frac{1}{K} \sum_{j=1}^K \left\{ \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} I(\tilde{y}_i^h = \hat{y}_i^m) \right\} \end{aligned}$$

where  $h \in \{1, 2, 3, 4, 5\}$

- Since notation is obviously extremely cumbersome, indices displayed will be limited only to those relevant.

### • Figure 1



- **Figure 1 Caption:** Training and Test set observation counts vs K
- We see that the possible training set observation counts using the CV method outlined above on the data provided range between 1247 and 2492 observations.
- The figure also shows that for K-values greater than 1248, test set sizes are confined to one.
- Training set sizes were selected for analysis by sampling 50 or more different values of K from k=5 to k=2490.
- Plots of comparing accuracies in the form of (2.1-1) are shown for each value of K selected as the value of training set length corresponding to K increases
- It was assumed that subsets taken from the data which were sampled randomly without replacement are at least partially representative of the population as a whole.
- It was assumed (incorrectly) for this analysis that test-set accuracy evaluations are comparable at all values of K.

## 2.3 Optimal Convergence Criterion

- The Pscore algorithm requires the specification of a confidence level for termination.
- For the purposes of this analysis, a confidence level of 0.75 was chosen.
- This value optimized probabilistic scoring classification accuracy of the first simulated validation outcome (see below) when the Psoring weights were trained on the full data set.

## 2.4 Simulated Validation Outcomes

### 2.4.1 Inferential Simulations

- Change traditional classification based upon the following assumptions:
  - Depression Classifications are a hierarchy, so outcomes for group 1 are lower than group 2,...etc
  - The 12 percent misclassification is due to the traditional scoring algorithm mis-classifying observations into a proximal class. Group 1 into group 2, group 3 into group 2 or group 2 into either group 1 or 2
  - Within each class there is a spectrum of conditions, and transitions between classes are essentially continuous.
- Based on the number of observations with the same traditional sum-score.
  - Let  $\mathbb{S}_T^{PR} = \left\{ i \mid S_i = T \right\}$  for  $i = 1, \dots, N$  and  $T = 0, 1, \dots, 27$  where  $S_i = \sum_{q=1}^9 a_q$
  - Then the number of observations with the same traditional sum-score ( $S_i$ ) is the value:  $|\mathbb{S}_T^{PR}|$
- Based upon distance from a threshold value. Meaning that observations which are closer to a threshold value are more likely to be mis-classified (switched)
  - If we suppose that  $\alpha_1 \leq \alpha_2 \in \{0, 1, \dots, 27\}$  are threshold values. The distance to the nearest threshold value of an observation  $i$  that is classified into traditional

sum value  $S_i$  is:

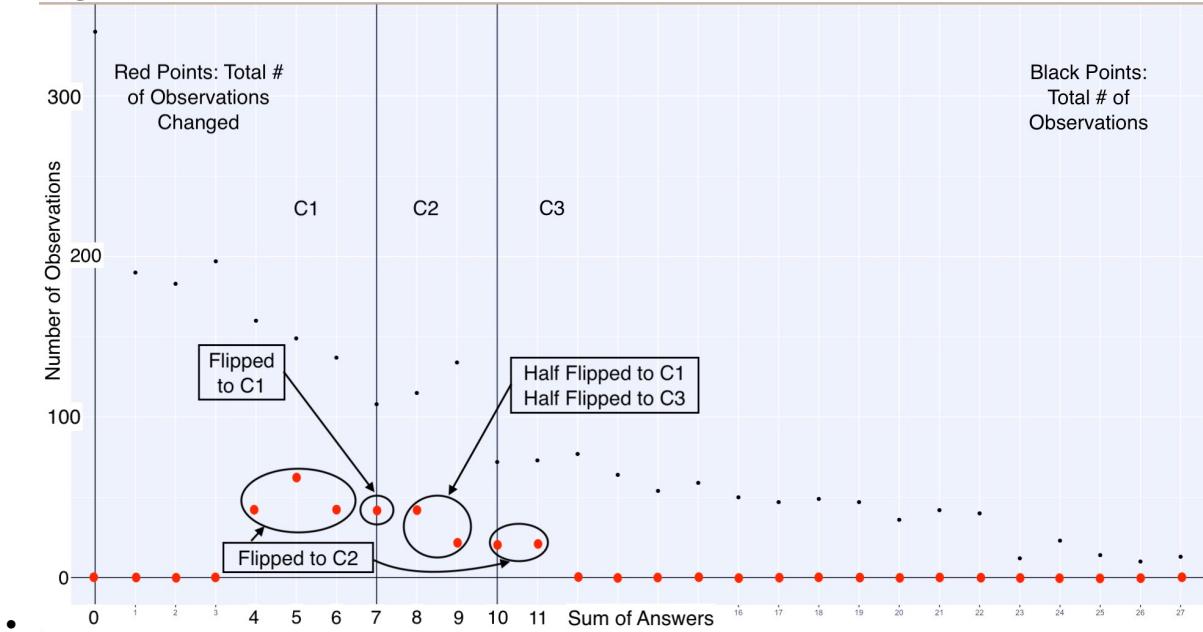
$$D_i = \min_{k=1,2} \left\{ \left| \alpha_k - S_i \right| \right\}$$

- The final probability of a traditional classification being altered in its representation in the simulated outcome data set then follows:

$$- P(C_i^{TR} \neq \tilde{C}_i^1) \propto \frac{|\mathbb{S}_T^{PR}|}{D_i}$$

– Proportionality constants were chosen so that the total difference between Traditional Scoring classification and simulated outcome was 12%

- **Figure 2:**



- **Figure 2 Caption:** Figure depicts observational distributions as classified by traditional methods, and how the first simulated validation outcome induced change
- Simulated validation outcomes generated using this method depended only on information relevant to the observation for which the outcome was being simulated
- Pairing of simulated validation outcomes performed with full-data
- Accuracy comparisons of Probabilistic Scoring and Traditional Scoring classifications were obtained on 100 different values of K using the CV algorithm previously outlined.
- The simulated outcome generated using this method was implemented in the process to obtain the optimal convergence confidence level in section 2.3

#### 2.4.2 Probabilistic Algorithm Simulations

- Motivating Principle: The Psoring algorithm incorporates more information contained in the data than anything else, therefore it should generate the most appropriate estimates
- Motivating Principle: The Psoring algorithm can estimate multiple amounts of inherently different estimates for the same observations based upon different training sets.

- The process:
  - Separate data into K-CV sets, and combine into train-test pairings as before in the CV analysis

$$(TR, TE)_j^K$$

where  $j = 1, \dots, K$

- For each training set, calculate the corresponding set of Pscore weights

$$TR_j^K \rightarrow \text{equation (1.3.2 - 1)} \rightarrow W_{AQj}^C$$

- \* K train-test pairs means there will be K distinct estimates of the training weights
- Calculate a representative weight set for the value of K by averaging over the K weight values

$$W_{AQk}^C = \frac{1}{K} \sum_{j=1}^K W_{AQj}^C$$

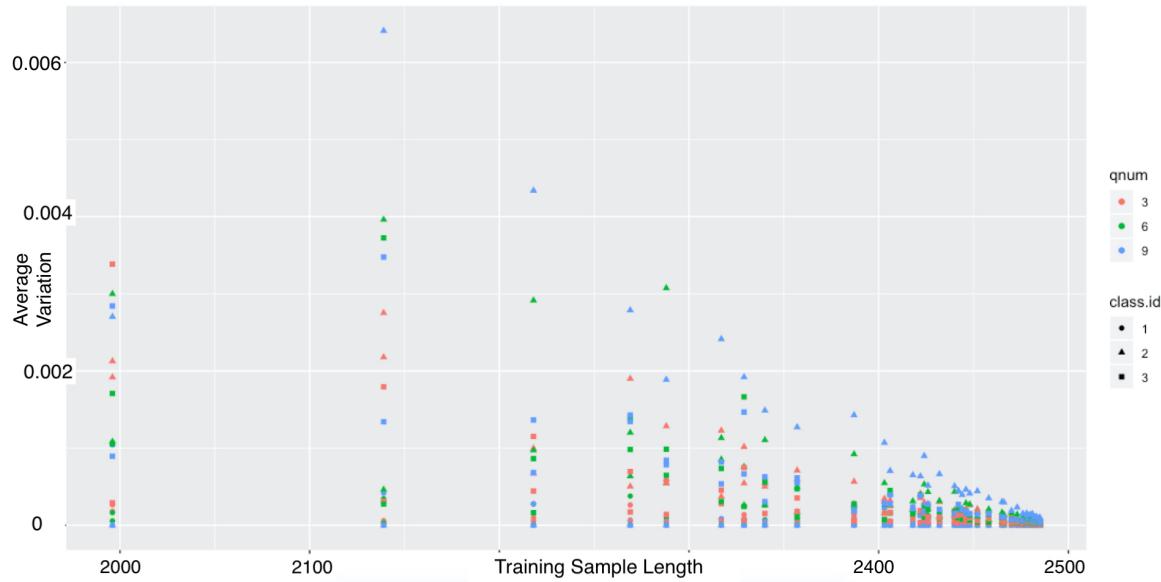
- For each of the K test sets use the calculated, representative weights to Probabilistically Classify a corresponding set of *simulated validation outcomes*

$$(W_{AQk}^C, TE_j^k) \rightarrow \text{equation (1.3.2 - 2)} \rightarrow \tilde{y}_j^2 k$$

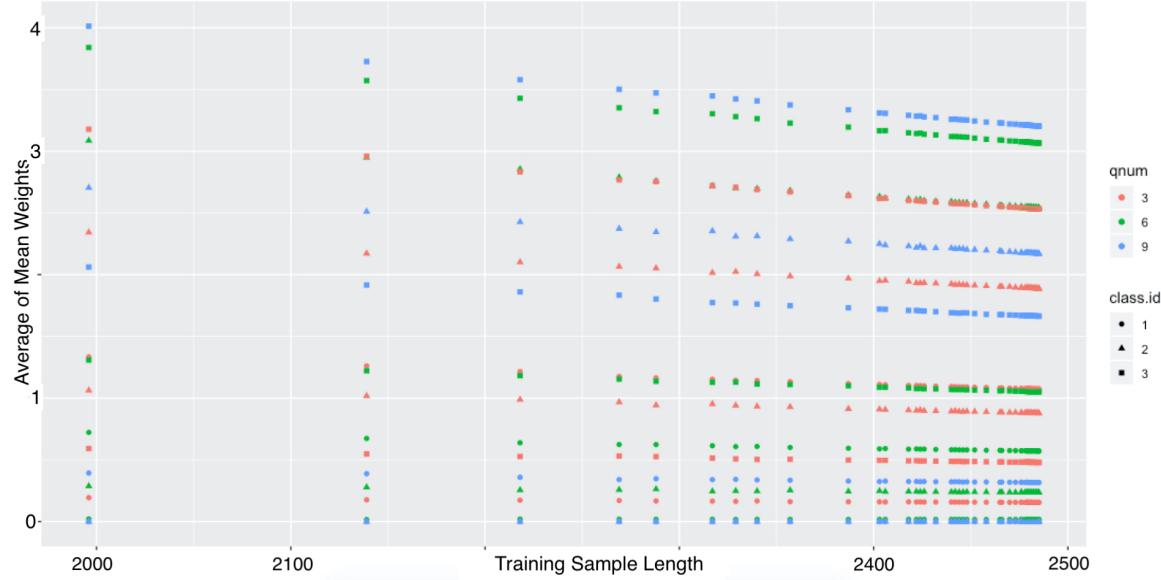
- The outcomes generated using this method depend on information obtained from a training set used to generate Pscore weights which were used in its classifications.
- Each simulated outcome will therefore be paired with data that was simulated using the same set of information. (i.e. categorized using the same training data).
- The values of the simulated outcome specific to training sample within each K value are then substituted into the CV algorithm for  $\tilde{y}^h$

$$\begin{aligned} \text{accuracy}_m^h (TR, TE)^k &= \frac{1}{K} \sum_{j=1}^K \left\{ \text{accuracy}_m^h (TR, TE)_j^k \right\} \\ &= \frac{1}{K} \sum_{j=1}^K \left\{ \frac{1}{|TE_j^k|} \sum_{i=1}^{|TE_j^k|} \left\{ I(\tilde{y}_i^h = \hat{y}_i^m) \right\} \right\} \end{aligned}$$

- The method makes the assumption that estimates of average weights are stable enough to produce reasonably consistent estimates
- **Figure 3:**



- **Figure 3 Caption:** Variance of average weight values for several question number/class ID combinations plotted against training data count length
- **Figure 4:**



- **Figure 4 Caption:** Mean value of the average weight estimates for several question number/class ID combinations plotted against training data count length
- Stability of the average estimate value, as training sample increases, in combination with a dramatic decrease in variance as training sample size increases are indications that the average weight values are capable of producing stable outcome estimates

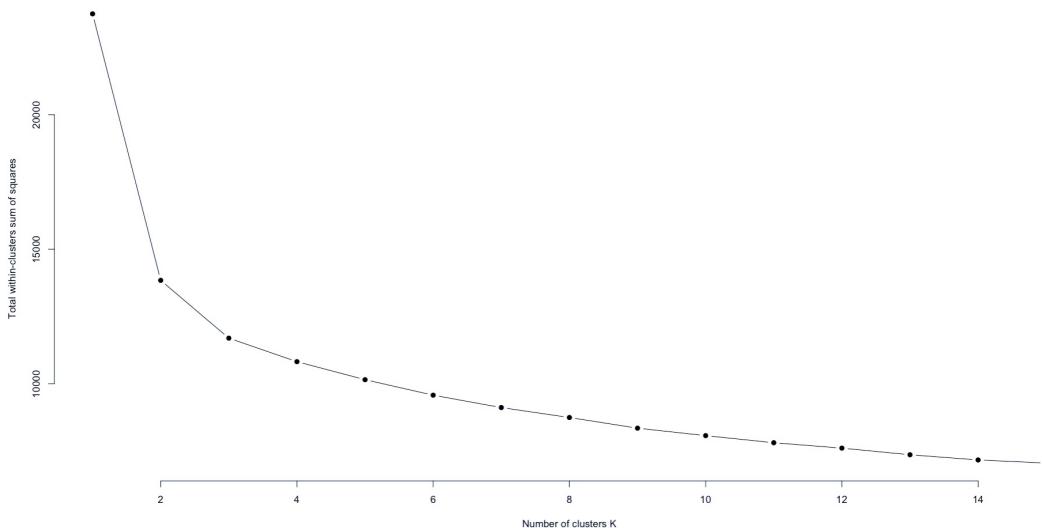
#### 2.4.3 Unsupervised Learning Classifications

- Using unsupervised learning techniques to deduce trends in the PHQ9 is a powerful tool

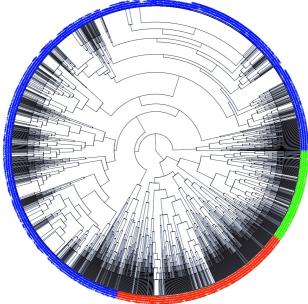
- The general algorithm we will exploit in the following fitting processes
  - Use an unsupervised method to cluster PHQ9 observations, preferably make as few assumptions as possible
  - Use the hierarchical aspects of the outcome classes (the fact that Depression classifications increase in magnitude) to group clusters into three clusters representing a simulated outcome classification for the observations contained within the cluster.
  - Input the simulated outcome classifications into the CV algorithm and obtain estimates for both Traditional Scoring and Probabilistic Scoring accuracies

#### 2.4.3.1 Kmeans and Heirarchical Clustering

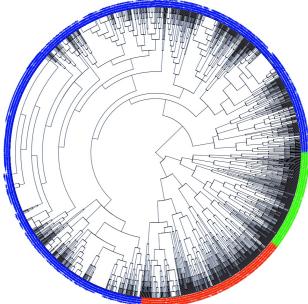
- Kmeans
  - The proper number of clusters in the Kmeans algorithm can be determined with the use of an elbow plot
  - **Figure 4:**



- **Figure 4 Caption:** A plot depicting the total within sum of squares vs the number of clusters chosen for the algorithm
- The elbow plot demonstrates that the data has a natural tendency to form three clusters, and therefore, no intermediate mapping is necessary.
- Hierarchical Clustering
  - A similar process was completed using two approached to Hierarchical clustering
  - The first distinguishes observations based upon Euclidean distance. A circular dendrogram, with corresponding classifications is displayed in **Figure 5** [5]
  - **Figure 5:**



- **Figure 5 Caption:** Circular dendrogram with three-category classifications
- The second hierarchical clustering technique distinguishes observations based upon Manhattan distance. A circular dendrogram, with corresponding classifications is displayed in **Figure 6** [5]
- **Figure 6:**

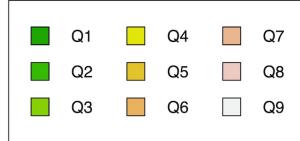


- **Figure 6 Caption:** Circular dendrogram with three-category classifications

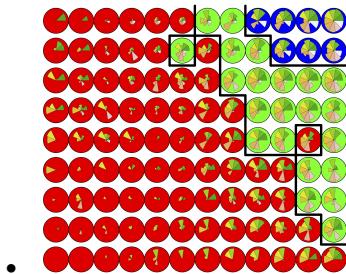
#### 2.4.3.2 Self Organizing Maps

- Self-organizing maps (SOMs) require a pre-specified structure (generally a lattice structure) to which observations are clustered based upon projected proximity
- Some advantages of using a SOM over Kmeans, or Principle Components Analysis are:
  - Non-linearity. The SOM algorithm has no underlying assumptions restricting model structure
  - Topological preservation. SOMs preserve ordering properties found in higher dimensions, when mapped to the projection space, therefore order (hierarchy) is preserved
  - Inferential potential. SOM visualization methods are extremely useful and meaningful.
- SOMs have standard pre-processing procedures that dictate standardization of observations prior to training.
- Three SOMs lattice structures were fit in this analysis. Each of these lattice structures was also fit with both the Euclidean and Manhattan metrics that determine the relationship between mapped observations in the projection space.
- The SOM lattice Structures implemented are:

- 12 X 9 rectangular topology
- 15 X 15 hexagonal topology
- 3 X 9 rectangular topology
- Plots of each of these mappings are found in **Figures -**
- **Figure 7:**



- **Figure 7 Caption:** Color coding key for question magnitude representation within each output lattice cluster
- **Figure 8:**

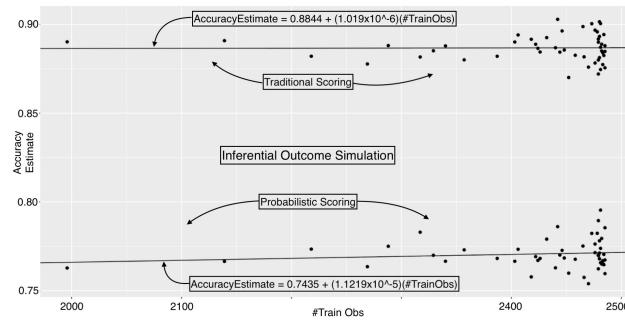


- **Figure 8 Caption:** Euclidean Norm, 12 X 9-rectangular topology SOM

### 3 Analysis and Results

#### 3.1 Inferential Outcome Simulation

- **Figure XX:**



- **Figure XX Caption:** displays estimated accuracy values for probabilistic scoring classifications and traditional scoring classifications across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

- **Table 1:**

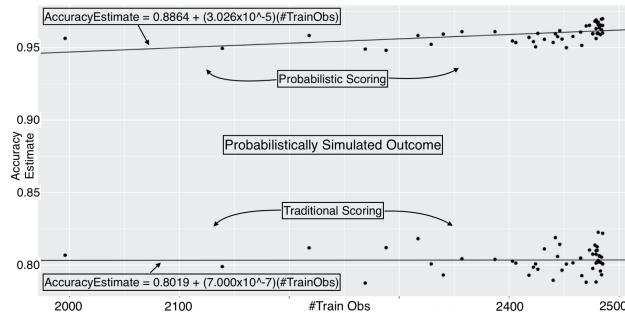
Parameter	Estimate	Std. Error	t. Value	Pr(> t )
Intercept	8.844e-1	2.902e-2	3.048e1	<2e-16
ID = Pscore	-1.409e-1	4.104e-2	-3.433e0	8.840e-4
#Train Obs	1.019e-6	1.197e-5	8.850e-2	9.323e-1
(ID = Pscore):(#Train Obs)	1.020e-5	1.692e-5	6.030e-1	5.481e-1

**Table 1 Caption:** summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

- After simulating an outcome using the Inferential Simulation method outlined in section (2.4.1) we compared the accuracy of probabilistic scoring and traditional scoring classification methods against this outcome for 50 different definitions of training sample length.
- The results of this comparison show that traditional scoring classification is more accurate than probabilistic scoring for all values simulated
- The results indicated that probabilistic scoring accuracy grew at a faster rate than traditional scoring accuracy; however these results lacked significance ( $p=0.5581$ )

## 3.2 Probabilistic Simulations

- **Figure XX:**



- **Figure XX Caption:** displays estimated accuracy values for probabilistic scoring classifications and traditional scoring classifications across 50 values of training set definitions. Also plotted are Linear Regression trend lines for each estimation method's accuracy estimates.

- **Table 1:**

Parameter	Estimate	Std. Error	t. Value	Pr(> t )
Intercept	8.864e-1	2.503e-2	3.541e1	<2e-16
ID = Traditional	-8.453e-2	3.540e-2	-2.388e0	1.891e-2
#Train Obs	3.026e-5	1.032e-5	2.932e0	4.210e-3
(ID = Traditional):(#Train Obs)	-2.956e-5	1.460e-5	-2.025e0	4.562e-2

**Table 1 Caption:** summary table for linear regression trend lines fit to accuracy estimate values displayed in **Figure XX**

---

## 4 Discussion

- Discussion of Results (separate from Results section)
  - Have you addressed your client's goals?
  - Are there any caveats to your analysis?
  - What are the next steps from such an analysis?
  - Recommended resources
- 

## 5 Appendix

- Necessary code or software instructions to carry out your analysis.
  - Include any resources you have referred to or found helpful
  - If appropriate, include other deliverables
- 

## 6 References

1. Kroenke K, Spitzer RL (2002) The phq-9: A new depression diagnostic and severity measure. *Psychiatric annals* 32: 509–515.
2. Kroenke K, Spitzer R (2010) Instruction manual: Instructions for patient health questionnaire (phq) and gad-7 measures.
3. Malik A More effective and cost effective use of the phq-9.
4. Kroenke K Spitzer, rl & williams, jb (2001). The phq-9. *Journal of General Internal Medicine* 16: 606–613.
5. (2020) *Sthda.com*.