Assignment #11 Text Mining XLMiner by Joshua Troup

Q1: Examine the term document matrix (TDM). Is it sparse or dense? Justify.

**Total # of 0s=2260**

**Total # of values greater than 0=740**

**The term document TDM proves to be sparse with the number of 0s exceeding the values greater than 0. No value exists for a given combination of dimension values.**

Q2: Find two nonzero entries and briefly interpret their meaning, in words (you do not

need to derive their calculation)**F79=3.68**

**F80=0**

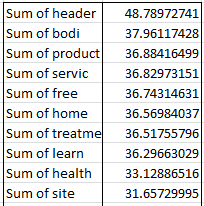
**The term “custom” has occurred in the document quite frequently cell F79 (3.68) however it does not appear frequent in the entire collection of the document. F78 also has an entry value of 3.68 in the “custom” term. This term shows up or is appears in the document 3.68 times.**

**The term “custom” has not occurred in the document F80 (0) but the value 0 appears more frequently in the entire collection of the document. No matching terms in the document for that particular row.**

Q3: Had you chosen the presence/absence option for the term document matrix, what

would those two nonzero entries be?**The two nonzero entries would have been 1’s. Of course, the 0 entries would remain 0’s.**

Q4: Present your table of Final List of Terms (Maximum 10 Top Terms)

****

Q5: Briefly explain the difference between the term document matrix(TDM) and the

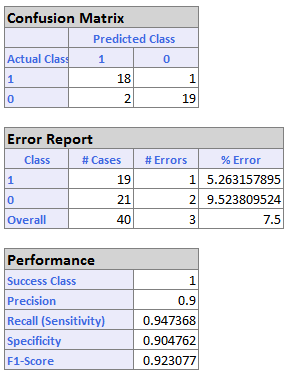
concept document matrix (CDM). Relate the CDM to what you learned in the principal

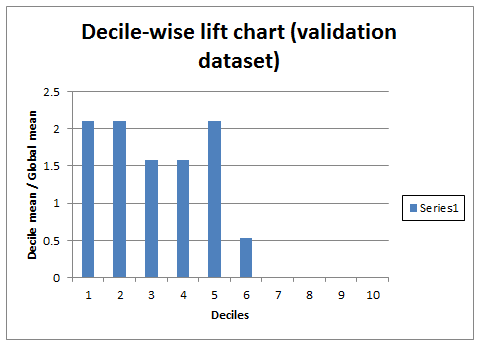
components (PCA) chapter (Chapter 4.)

**Term Document Matrix (TDM) and Concept Document Matric (CDM) perform similar tasks. The difference is TDM is focused on each single term analyze and showing exact cell and value. CDM join similar terms into concepts which are the columns. The documents are listed down the left side of the matrix. Principle Component Analysis (PCA) relates to CDM in the way of both algorithms producing concepts or combinations of correlated values. PCA then creates new predictor variables. CDM uses latent semantic indexing (LSI) which uses combined terms to produce concepts. The effect is reduction in dimension of data.**

Q6: Using logistic regression, partition the data (60% training, 40% validation), and develop a

model to classify the documents as 1’s or 0’s. Comment on its efficacy. (sample data partition

settings shown below)****

****

**19 classes were classified correctly for class 1 with 1 error resulting in a 5.26% Error.**

**21 classes were classified correctly for class 0 with 2 errors resulting in a 9.52% Error.**

**Overall, 40 classes were classified correctly with 3 errors. 7.5% Error. The decile lift chart is not as relevant as the LR Output since the goal is to classify all ads, not to select which ads are most likely relevant.**

Q7: Why use the concept document matrix(CDM), and not the term document matrix

(TDM), to provide the predictor variables?

**The answer is very similar to Q5, CDM is used for predictor variables to reduce dimension in data. Grouping similar terms into concepts helps this process and develops a well suited model.**