Homework 5

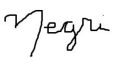
Group 3
Nathalia Negri
Troy Yang

305-318-3446
781-812-4931

negri.n@husky.neu.edu
yang.tr@husky.neu.edu

Percentage of Effort Contributed by Student 1:_____50%_____

Percentage of Effort Contributed by Student 2:_____50%_____

Signature of Student 1:_____  **_____**

Signature of Student 2:_____  **_____**

Submission Date:_____June 22, 2017_____

Question 11.3
a. 30 Epochs:

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
918782292.3	1032.411	326.0895

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
730182550.8	1127.872	259.1502

300 Epochs:

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
358732924.9	645.1073	144.9275

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
863612366.8	1226.601	92.38265

3000 Epochs:

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
48067218.67	236.1407	21.5638

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
2122993807	1923.173	-66.3665

10000 Epochs

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9977477.776	107.5862	18.85944

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3169745799	2349.937	-100.871

i. The RMS error for the training data decreases as the number of epochs increases.

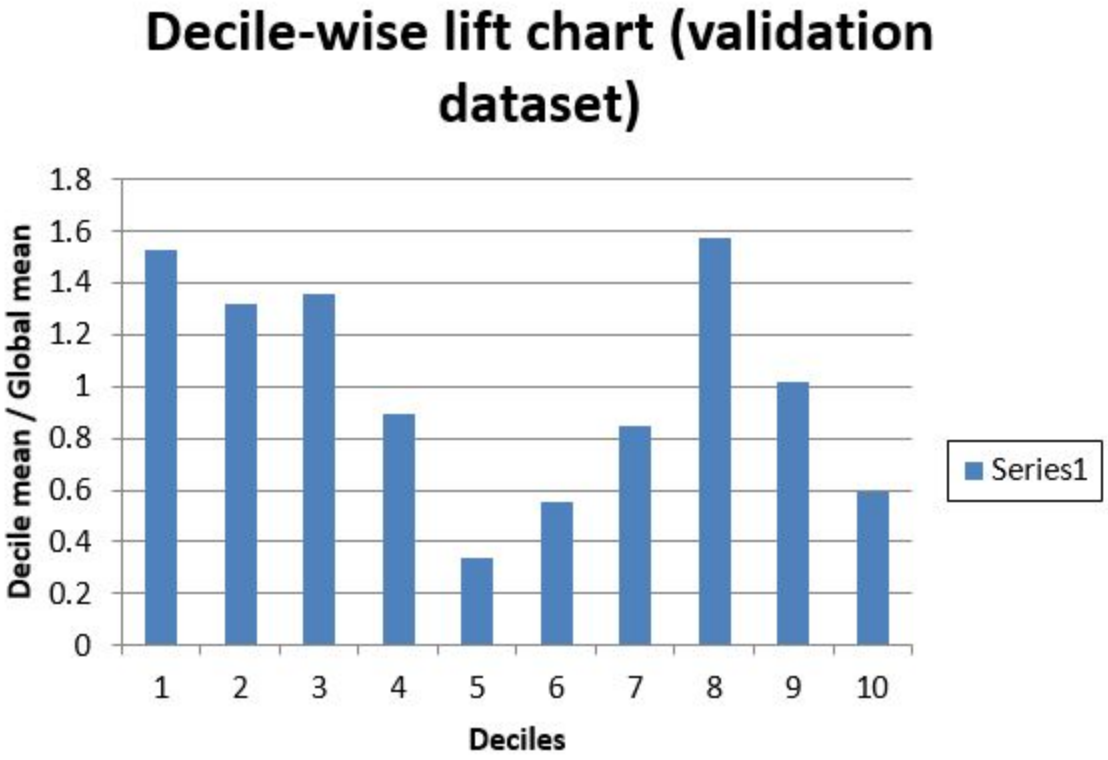
ii. The RMS error for the validation data increases as the number of epochs increases.

iii. As the number of epochs increases the model begins to overfit the training data. An epoch is an iteration of updating the weights in the model. So, after too many iterations, the weights begin to exactly fit the training data. Based on the validation data error reports it appears the appropriate number of epochs is 30.

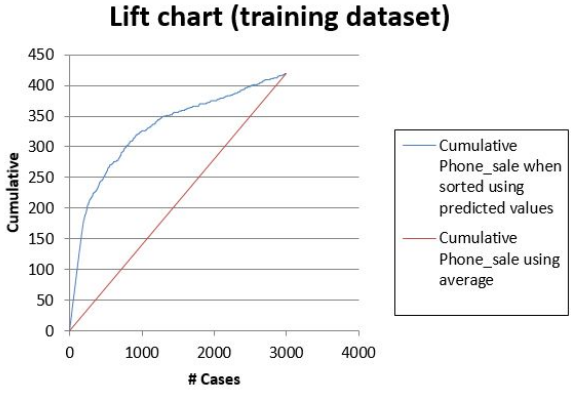
Question 11.4

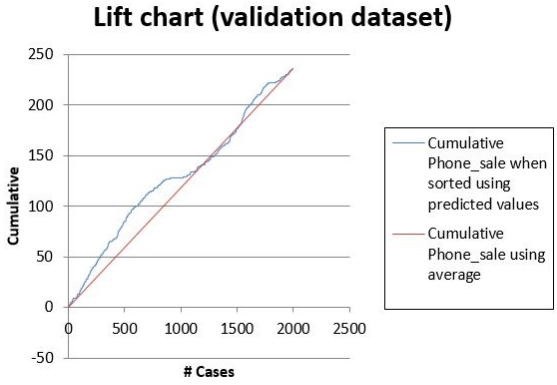
a. The leftmost bar in the decile-wise lift chart shows the relationship between how the

model performs versus the average. According to this decile-wise lift chart, in the first 10% of new customers the neural network model is expected to result in about 150% more phone service sales than if the average were used.

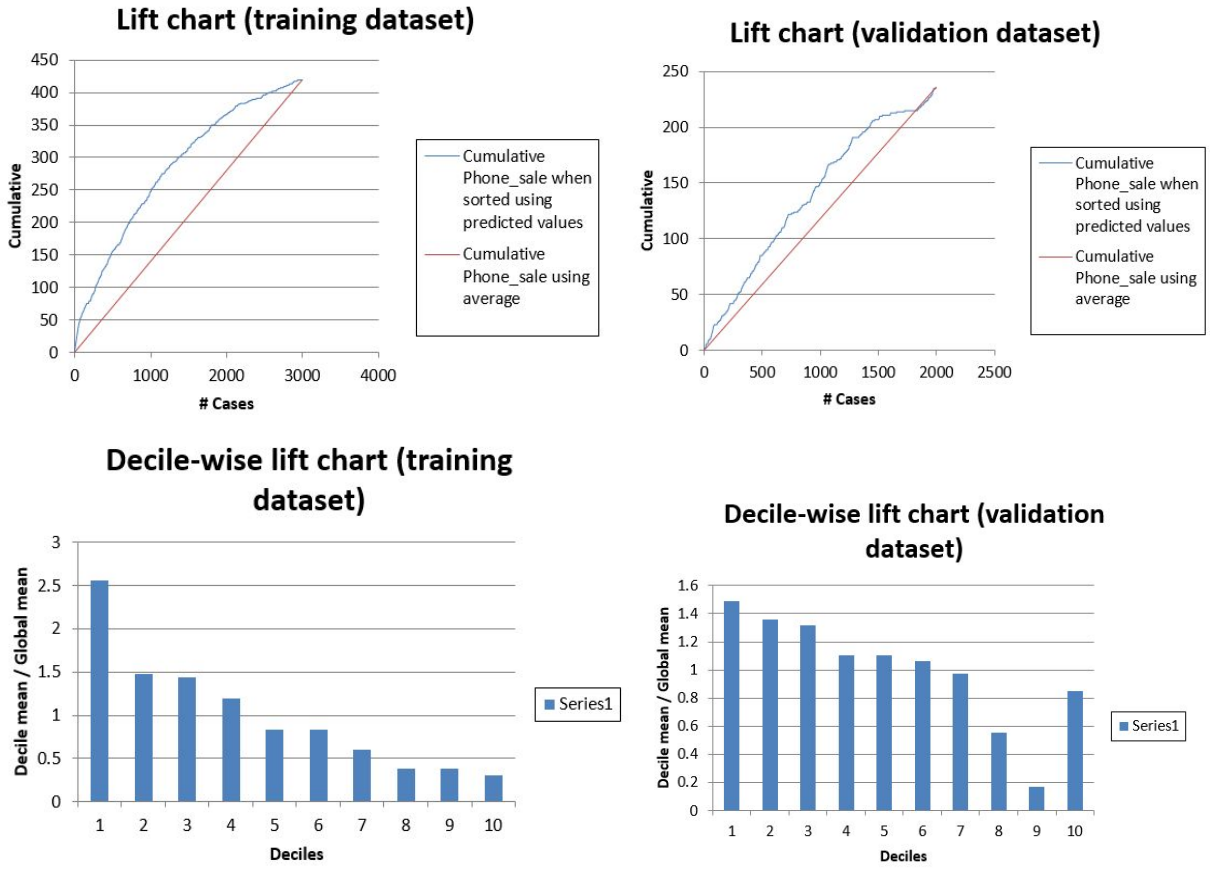


b. As shown below, one can clearly see that the training data lift chart looks much better than the validation data lift chart. The reason for this is because the training data itself was used to create the neural network, and because the number of epochs was very high, the model was most likely overfit on the training data, which explains both the great performance on the training data, as well as the relatively abysmal performance on the validation data.





c. Aside from the first and tenth deciles, the lift charts for the validation and the training data look pretty similar. As stated in (b), 3000 epochs is a lot, and it was very likely that the neural network was overfit on the training data in that case. However, we only selected 100 epochs, so this model is probably much more generalized, and as a results performs much better on the validation data.



d. Although we can view the inter layer connection weights in XL Miner, these weights do not provide us with any significant information about the effects of the various variables. This is one common criticism of neural networks, as it is somewhat comparable to a black box when compared to models like multiple linear regression.

Question 12.3

a. W_3d, remove, internet, receive, addresses, free, business, credit, money, C$, and CAP_long appear to be the 11 predictors that vary the most between spam and not-spam. We determined this by examining the colMeans in Excel.

b.

XLMiner : Discriminant Analysis

Date: 21-Jun-2017 18:35:35

Output Navigator				
Inputs	Prior Class Probabilit	Train. Score - LDA S	Valid. Score - LDA S	LDA Train. Lift Chart
LDA Train. Detail Rpt	LDA valid. Lift Chart	LDA Valid. Detail Rpt		

Elapsed Times in Milliseconds			
Reading Data	Computatio	Writing Dat	Total
13	20	50	83

Data	
Workbook	12.3_Spambase (1).xlsx
Worksheet	Data_Partition
Training data used for building the model	$B$21:$M$2781
# Records in the training data	2761
Validation data	$B$2782:$M$4621
# Records in the validation data	1840

Variables											
# Input Variables	11										
Input variables	W_3d	remove	internet	receive	addresses	free	business	credit	money	C$	CAP_long
Output variable	Spam										

Parameters/Options	
Use Linear Discriminant Analysis	Yes
Use Canonical Variate Analysis	No

c.

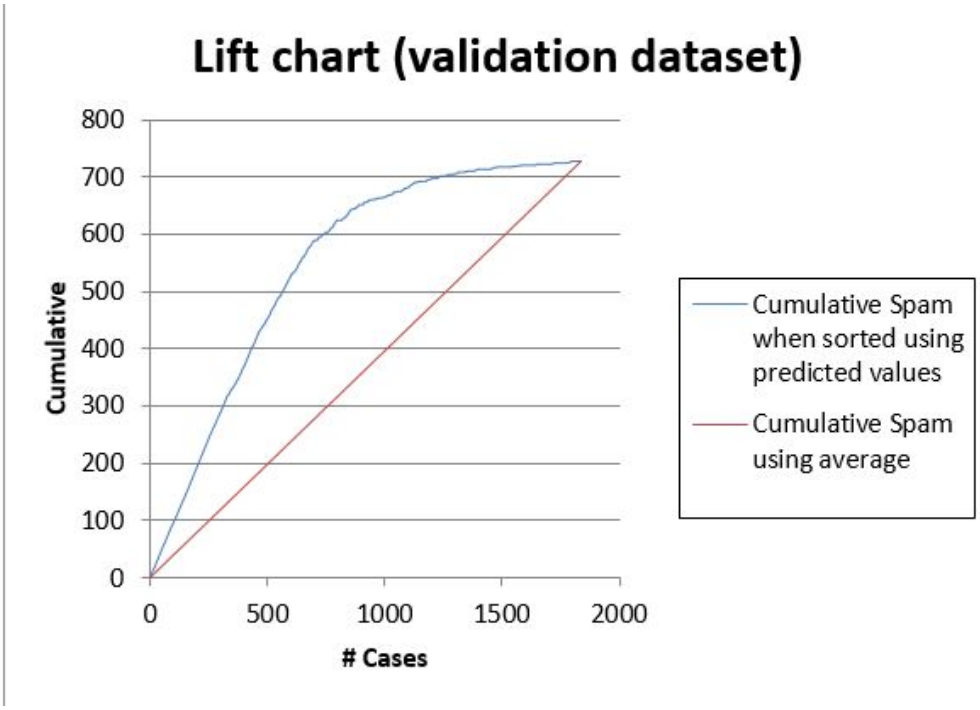
Validation Data LDA Scoring - Summary Report

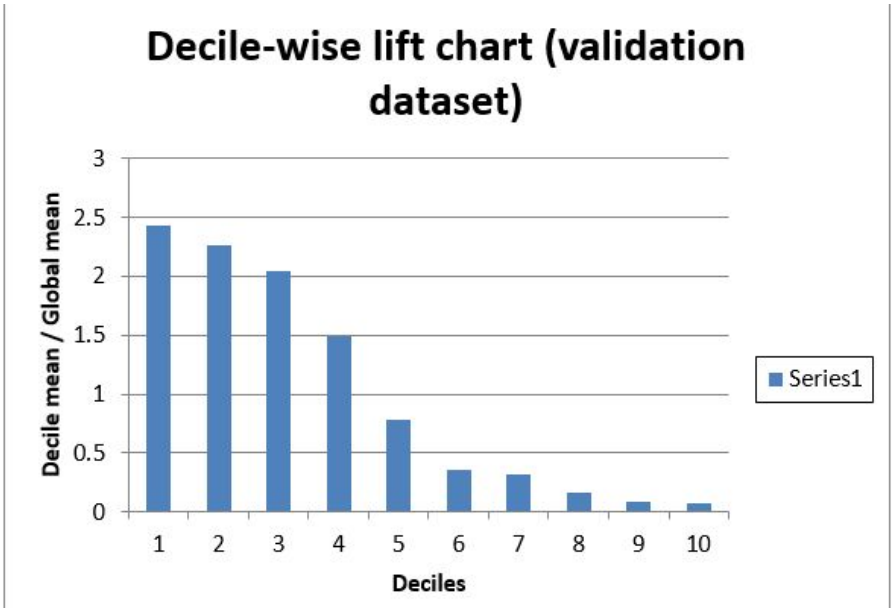
Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	404	325
0	32	1079

Error Report			
Class	# Cases	# Errors	% Error
1	729	325	44.58162
0	1111	32	2.880288
Overall	1840	357	19.40217

Performance	
Success Class	1
Precision	0.926606
Recall (Sensitivity)	0.554184
Specificity	0.971197
F1-Score	0.693562





Based on the three figures above, it would appear that the models are quite good. That is true for non-spam emails, this model has around a 45% error rate when it comes to spam emails, which it make it not very useful.

d. Original Spam Const. = -2.3355
Original Non Spam Const. = -0.5211

New Spam Const. = -2.3355 + log(0.10) = -3.3355
New Non Spam Const. = -0.5211 + log(0.90) = -.56686

e. Spam Ratio = 1
Non Spam ratio = 20
New Non Spam Constant = Original Non-spam Const. + log(Non Spam ratio/Spam Ratio)
New Non Spam Const, = -0.5211 + log(20) = 0.7799

New Spam Constant = Original Spam Constant = -2.3355

Question 14.3

a. In Transaction 12, the consumer bought nail polish, brushes, concealer, and bronzer. In Transaction 8, the consumer bought nail polish, brushes and bronzer. In transaction 4, the consumer bought nail polish, brushes, concealer and bronzer. In transaction 1, the consumer bought blush, nail polish, brushes, concealer and bronzer.

b.
i. Confidence % = 80.52 = 62/77 where 62 is the support for A and C and 77 is the support for A

ii. Support for A is the number of transactions that include brushes and concealers, which is 77, and support for C is the number of transactions that include nail polish and bronzer, and support for A and C is the number of transactions that include all of the previously stated items.

iii. The lift ratio is the confidence of the model divided by the benchmark confidence (just the average). It is an indicator of how much better or worse your model performs vs. using the average. The larger the lift ratio the greater the strength of the association.

iv.
If the customer bought brushes and concealer than they will also buy nail polish and bronzer.

v.

XLMiner : Association Rules

Output Navigator	
Inputs	List of Rules

Elapsed Times in Milliseconds		
AssocRules Time	Report Time	Total
49	3	52

Inputs

Data	
# Transactions in Input Data	1000
# Columns in Input Data	14
# Items in Input Data	14
# Association Rules	139
Minimum Support	100
Minimum Confidence	50.00%

List of Rules

Rule: If all Antecedent items are purchased, then with Confidence percentage Consequent items will also be purchased.

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	100	Brushes	Nail Polish	149	280	149	3.571428571
2	53.21428571	Nail Polish	Brushes	280	149	149	3.571428571
3	65.14285714	Mascara & Eyeliner	Concealer & Eye shadow	175	201	114	3.240938166
4	56.71641791	Concealer & Eye shadow	Mascara & Eyeliner	201	175	114	3.240938166
5	64.67391304	Blush & Mascara	Concealer & Eye shadow	184	201	119	3.217607614
6	59.2039801	Concealer & Eye shadow	Blush & Mascara	201	184	119	3.217607614
7	65.38461538	Blush & Eye shadow	Concealer & Mascara	182	204	119	3.205128205

vi.

Row ID	Confidence %	Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Lift Ratio
1	100	Brushes	Nail Polish	149	280	149	3.571428571
2	53.21428571	Nail Polish	Brushes	280	149	149	3.571428571
3	65.14285714	Mascara & Eyeliner	Concealer & Eye shadow	175	201	114	3.240938166
4	56.71641791	Concealer & Eye shadow	Mascara & Eyeliner	201	175	114	3.240938166
5	64.67391304	Blush & Mascara	Concealer & Eye shadow	184	201	119	3.217607614
6	59.2039801	Concealer & Eye shadow	Blush & Mascara	201	184	119	3.217607614

Rule 1: If they buy brushes, then they'll buy nail polish;
Rule 2: If they buy nail polish, then they'll buy brushes;
Rule 3: If they buy mascara & eyeliner, then they'll buy concealer & eye shadow.

vii. The first dozen couples are basically conjugate pairs, thus creating a lot of redundancy. In terms of their utility, you can just use the one from the pair that has the higher confidence.