MIS 432 – Project 1 – Spring 2020

The dataset is *Tayco.cvs* and can be found in our Blackboard course.

Tayko is a software catalog firm that sells games and educational software. It recently put together a revised collection of items in a new catalog, which it is preparing to roll out in a mass mailing. A sample mailing was sent out to test the market. After using this dataset of customers, they want to be able to predict who will be a purchaser from their catalog in order to target the mass mailing. The following table explains the variables relevant to this project.

Variable Name	Response or Predictor	Description	Variable Type	Code Description
US	Predictor	Is it a US address?	Binary	1: Yes 0: No
Freq	Predictor	Number of transactions in last year	Numerical	
last_update_days_ago	Predictor	How many days ago last update was made to customer record	Numerical	
1st_update_days_ago	Predictor	How many days ago first update to customer record was made	Numerical	
Web_order	Predictor	Customer placed at least one order via web	Binary	1: Yes 0: No
Gender	Predictor	Customer is male	Binary	1: Yes 0: No
Address_is_res	Predictor	Address is a residence	Binary	1: Yes 0: No
Purchase	Response	Person made purchase from catalog	Binary	1: Yes 0: No

Case Study Dataset is from Data Mining for Business Analytics: Concepts, Techniques, and Applications in R, 2017 edition, by Shmueli, Bruce, Yahav, Paten, and Lichtendahl, Wiley Publications

Project Description: The objective of this project is to evaluate and compare three neural network models that differ by how each hidden layer structure is defined. Explore which of the following three hidden layer structures performs the best according to their accuracy percentages.

1) one hidden layer with two nodes

2) one hidden layer with three nodes

3) two hidden layers, each with two nodes.

Project Requirements:

1. First, pre-process the dataset so that the three numerical variables are each scaled by:
$$\frac{X-min}{max-min}$$
The max and min should be calculated for each numerical variable separately.
Also, expand the response variable into two columns: **Purchase_No** and **Purchase_Yes**, where Purchase_Yes is 1, if Purchase=1, and Purchase_No is 1, if Purchase = 0. Don't erase the original Purchase variable, since it will be used in the Confusion Matrix. This pre-processing step can be done in Excel and then saved as a .csv file.
2. Partition the dataset into 60% training and 40% validation datasets. Use set.seed(1)
3. Perform the three neural network fittings each on the same training dataset. The three models differ only by the hidden structures as stated above in the Project Description. Predict the response variables: **Puchase_No** and **Purchase_Yes.** Use seven predictor variables which include the three scaled variables and the four binary predictors listed in the table above.
4. Plot each of the three neural networks.
5. Construct a Confusion Matrix on the validation dataset for each of the three neural network fittings.
6. Compare the Confusion Matrices through their error and accuracy percentages.
7. In conclusion, write up a recommendation for the best of the three hidden layer structures for Tayko.

Project Report and Submission

1. Write up your submission in a word document report explaining the process and your results. Include graphs of the networks and the important outputs from R within the text of your report.
2. Include an appendix at the end of all of your R code and outputs copied and pasted from the Console window.
3. The length of the report should be at least 800 words (not including the outputs or the appendix).
4. Save the word file as "Last Name First Initial Project 1" and submit in the Assignments/Projects link in Blackboard.

Case Study Dataset is from Data Mining for Business Analytics: Concepts, Techniques, and Applications in R, 2017 edition, by Shmueli, Bruce, Yahav, Paten, and Lichtendahl, Wiley Publications