Week 6 Assignment

Note: Late homework assignments are not accepted because the solutions are discussed at the start of every class on Tuesday. Your solution to this assignment must be uploaded on eCampus as a PDF file. A common approach is to put the solution into a word document and then save that into a PDF file. Please only submit PDF files.

Assignment: You can complete one of the two parts to this assignment, or both. If you do both parts, you can obtain extra points for this assignment. You are expected to complete at least one part successfully.

Data File: CreditHistory_Clean.xlsx

Part 1: Create a SAS EM project names "Week 5 Homework". In that project read this data file for this assignment. Import the data, ensuring that all attributes have the proper metadata described in the data dictionary. In this case, the target is 'good_bad', a binary target.

Do not change any of the random seeds (default values of 12345).

These data do no contain missing values or outliers. There is no need to "clean" these data. However, please ensure the interval variables are all standardized. This is a default setting in SAS EM.

Also, use 30 initial starting points (run) with a maximum or 20 iterations each.

If you are using the non-HP neural network, use 'logistic' *activation f*or your hidden layer perceptrons, and one-layer networks with 3, 5, 7, 9 and 11 perceptrons.

If you are using the HP Neural Network, there is no option for logistic activation. Just use the default in that case. For network configurations use the following five configurations:

1. One hidden layer with 3 perceptrons
2. One hidden layer with 11 perceptrons

3. Two hidden layers, the first with 5 perceptrons and the second with 4.
4. Two hidden layers, the first with 6 perceptrons and the second with 5.
5. Two hidden layers, the first with 7 perceptrons and the second with 6.

Use 10-fold cross validation to determine the best neural network. For each setting of the parameter, calculate MISC, recall, accuracy, precision and F1. Use these metrics to select the best number of perceptrons for your network, based upon the 10-fold cross-validation.

You will need to calculate these metrics by hand, outside of SAS EM. The easiest way is to copy the confusion matrix from SAS EM into Excel for each of the five different network configurations. You can find the confusion matrix by opening the results window for each network configuration and clicking on:
"view"->"assessment"->"event statistics matrix"

You can copy and past that window directly into Excel. I would add an additional column labeled "perceptrons" and enter the number of perceptrons for that matrix. Finally you will need to enter the excel formulas for calculating the evaluation metrics for each of the 5 networks.

After you have selected the best number of perceptrons, evaluate your model using a 70/30 training/validation split. Calculate the same metrics for the validation data in this fit.

Part 1: SAS EM Solution Upload (all screen shots must be readable)

1. A screen shot of your project diagram
2. A screen shot or listing of ALL SAS code used in your diagram.
3. A table of the metrics for each of the 10 cross-validation folds
4. Describe which model you selected and why.
5. A table of the same metric for the 70/30 test of your selected model.
6. Compare the best neural network you found against the best tree you found in the week 5 assignment.

Part 2: Do the same assignment as Part 1 using Python.

Use one-hot encoding for the nominal attributes, and scale the interval attributes using 'std' (standardized scaling). Evaluate the following 5 network configurations:

6. One hidden layer with 3 perceptrons
7. One hidden layer with 11 perceptrons
8. Two hidden layers, the first with 5 perceptrons and the second with 4.
9. Two hidden layers, the first with 6 perceptrons and the second with 5.
10. Two hidden layers, the first with 7 perceptrons and the second with 6.

Use 'logistic' activation for all networks, and random_state=12345.

Note that in these python neural networks, you cannot have direct connections from the inputs to the output node(s). As a result your best network will not be the same as that found by SAS EM.

Running cross validation with Neural Networks can take some time on your computer. Reuse your homework 5 solution. Simply replace the decision tree nodes with neural network nodes, and it should run fine, but slower.

Please download the latest version of *Class_replace_impute_encode* and *Class_FNN*. Also use the parameter *drop=False* with the *ReplaceImputeEncode()* method so that the last one-hot column is not discarded. Dropping the last column is only needed with10 linear and logistic regression.

From *Class_FNN* you will find some methods for displaying the results from these analyses.

Prepare a report containing:

1. A listing of your python code.

2. A table of the metrics (recall, accuracy, precision and F1) calculated for each of your 10 cross-validation folds.
3. Describe which model you selected from Cross-Validation, and why.
4. A table of the metrics (recall, accuracy, precision and F1) for the 70/30 split using your selected model.
5. Compare the best neural network you found against the best decision tree you found in the week 5 assignment.