Multi-Layer Perceptron Neural Net Model Identifies Effect in Random Data

Ariel Linden, Dr.P.H. and Paul R. Yarnold, Ph.D.
Linden Consulting Group, LLC  
Optimal Data Analysis, LLC

Prior research contrasted the ability of different classification algorithms [logistic regression (LR), random forests (RF), boosted regression (BR), support vector machines (SVM), classification tree analysis (CTA)] to correctly fail to identify a relationship between a binary class (dependent) variable and ten randomly generated attributes (covariates): only CTA found no relationship. In this paper, using the same ten-variable N=1,000 dataset, a Weka™ multi-layer perceptron (MLP) neural net model using its default tuning parameters yielded (area under the curve) AUC=0.724 in training analysis, and AUC=0.507 in ten-fold cross-validation. With the exception of CTA, all machine-learning algorithms assessed thus far have identified training effects in completely random data.

Predictive accuracy of CTA was compared with accuracy obtained using LR, RF, BR and SVM classification algorithms. In that research, an artificial dataset was created with 500 “group 1” and 500 “group 2” observations as well as ten randomly created continuous covariates (attributes) which by design have no association with the binary dependent (class) variable. Among all algorithms tested only CTA correctly failed to discriminate among random groups.

This study assesses whether a consistent finding occurs for models identified by MLP analysis using the default tuning parameters: learning rate 0.3, momentum 0.2, 500 training epochs, and hidden neurons equal to (10 inputs and 2 classes/2).

As done previously, receiver operating characteristics (ROC) analysis was conducted treating actual class status as the reference variable, and treating predicted probabilities from the MLP model as the classification variable. A model that perfectly discriminates two groups has AUC=1.0, and a model that provides chance-level discrimination has AUC=0.50.

Results obtained by MLP in training analysis are summarized in Table 1: CTA was unable to identify a model so reproducibility analysis wasn’t possible. The corresponding effect strength for sensitivity or ESS index (on which 0 is the classification accuracy expected by chance, and 100 is perfect accuracy) is ESS=39.4, indicating an effect of moderate strength.

Table 1: MLP Model Training Results

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Group 1</td>
<td>322</td>
<td>178</td>
<td>64.4</td>
</tr>
<tr>
<td>Class Group 2</td>
<td>125</td>
<td>375</td>
<td>75.0</td>
</tr>
<tr>
<td>Predictive Value</td>
<td>72.0</td>
<td>67.8</td>
<td></td>
</tr>
</tbody>
</table>
Ten-fold cross-validation analysis results obtained for the MLP model are given in Table 2 (ESS=0.8, a very-near-chance result).

Table 2: MLP Model 10-Fold Results

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Group 1</td>
<td>272</td>
<td>228</td>
<td>54.4</td>
</tr>
<tr>
<td>Class Group 2</td>
<td>268</td>
<td>232</td>
<td>46.4</td>
</tr>
<tr>
<td>Predictive Value</td>
<td>50.4</td>
<td>50.4</td>
<td></td>
</tr>
</tbody>
</table>

The failure of the MLP model to replicate in cross-validation reasserts the need to conduct reproducibility analysis and supports general constraint of CTA models to only retain attributes having stable effects in training and LOO analysis within the model.5-7

This paper supports the results obtained in previous analysis3,5 identifying an important limitation of machine learning algorithms often used for predicting binary outcomes (e.g., to obtain propensity scores8-11). That is, an MLP model is likely to find relationships between variables which really don’t exist. These results should be independently replicated, and the limits of this phenomenon should be identified. Future research should assess the effect of the number of random attributes available to the algorithms, the number of significant digits used for measures (an index of precision of measurement), and the number of class category levels in the application, in affecting the training and validity AUC. Research should also investigate applications involving randomized categorical attributes having different numbers of levels.

Finally, the findings continue to support our recommendation to employ the ODA and CTA frameworks to draw causal inferences regarding treatment effects in observational data, and in data from randomized controlled trials.8-24 A large, ever-increasing body of evidence supports the use of ODA and CTA to evaluate the efficacy of health-improvement interventions and policy initiatives.25-27

References


**Author Notes**

No conflict of interest was reported.