Challenge 21 Report

Robert Erick

August 16, 2023

**Overview:**

The purpose of the challenge is to create a model (as accurately as possible) from a charity contribution revenue data set of about 11 columns. The goal is to determine what factors would predict the receipt of charity contribution revenue.

The methodology is a follows:

* Drop 2 inconsequential columns from the data
* For 2 of the data columns (application\_type and classification), bin the least significant data into “other” bins
* Replace the categorical data with integer encoded data
* Scale the data
* Divide the data into train/test groups
* Create a deep learning model of several layers by guessing about the various parameters and architecture
* Build and fit the model and then assess the accuracy of the model
* Save the model to an hdf5 file

After that is all done, I was to circle back around and try to optimize the model further.

**Results:**

The results of the first build of the model are less than stellar. The model indicates an accuracy of 72.60%.

The results for the optimization were not much better. I did the following things:

* Binned the application\_type and classification columns more severely, creating a function that would do so
* I used some code to drop highly correlated columns
* Encoded categorical data
* Kept only the top 15 most important columns
* Loop over different hyper parameters
  + Ran the keras tuner for various changes to hyper parameters
  + Parameters like a) number of hidden layers b) differing numbers of nodes c) different optimizers d) different learning rates e) different activation functions
* It was important to delete the old results folders used by the tuner in order to run correctly and repetitively

All of that resulted in a model with fewer columns and data but with an accuracy of 72.12%. I couldn’t get the model to improve under optimization.

It was interesting, however, that a model of 68.80% accuracy could be had by just keeping the 3 most important columns. Those columns are: affiliation\_companysponsored, affiliation\_independent, and organization\_association. That simple model used activation=relu, optimizer=adam, learning rate = 0.01, 2 layers, 2 epochs. Those settings were determined by a 20 minute keras tuner evaluation. All other columns aside from the 3 listed were removed from the 68% simple model.

**Summary**

Regarding the first model, I would conclude that its accuracy was difficult to beat an I could only equal it even by making use of keras tuner and further adjustments.

Regarding the second, optimized model, I could achieve 72% if I left about 10 most significant columns. I could achieve 68% if I chose only the 3 most significant columns.

I think the insight gained from the 68% model is that company sponsorships, independent affiliation.

In fact, I also trained a model for 68% accuracy which had only 1 column: affiliation\_company\_sponsored. So…if a company sponsors a charity event, success for giving is much increased and that is the dominating factor.