Neural Network Training Using Tensor Flow

Name: Rudy Duvnjak

Date: 02/17/2022

Overview

Objective of this machine learning exercise it to be able to predict success od AlphabetSoup Charity donation with accuracy higher than 75% based on data collected in charity_data.csv file. Tensor flow was used as a primary tool to train data, and make prediction accuracy. During the data preparation process Standard Scaler was used to optimize data further. In addition, binning was performed completed on APPLICATION_TYPE, CLASSIFICATION, and NAME columns (in the optimized version). For enumeration OneHotEncoder method was used.

Data Processing

- Target variable used was "IS_SUCCESSFUL"
- Feature variables were as follows:
  - "NAME"
  - "APPLICATION_TYPE"
  - "AFFILIATION"
  - "CLASSIFICATION"
  - "USE_CASE"
  - "ORGANIZATION"
  - "STATUS"
  - "INCOME_AMT"
  - "SPECIAL_CONSIDERATION"
  - "ASK_AMT"
  - "NAME" (NAME was used only in optimized model)
- In the initial model "NAME" and "EIN" were dropped as insignificant variables
- "NAME" variable was binned and used in optimized model

Compiling, Training and Evaluating Model

- 3 layers were used with 120, 40, 10 and 5 neurons respectively, plus 2 ReLU and 1 Sigmoid activations were used. Primary reason of using 4 layers was to ensure that data is not underfitted
- Number of Neurons in the layers was determined by number of features in enumerated dataframe. We had about 250 features, so reducing number of neurons to 120 in the first hidden layer seamed appropriate, and then reduced following layers accordingly

- As seen in the optimized model notebook, 77.8% accuracy was achieved which met expected criteria of 75% or higher
- Step used to increase accuracy was keeping NAMES column and binning all names with count of less than 10 into Other category under names. Increasing number of layers and neurons due to number of features helped improve accuracy of the model as well

Summary

In summary, over model accuracy improved substantially after NAMES column inclusion, increase in number of hidden layers. So decision was made to use 2 ReLU activations, and use of Sigmoid in third hidden layer and output layer. As mentioned above accuracy improved in optimized model to about 78%, which satisfied our criteria for the model. Considering that we had IS_SUCCESSFUL column as a target we could have used supervised learning as well to model accuracy.