Deep Learning Challenge Report

Overview

The goal of this challenge was to assist a foundation in selecting applicants for funding. We needed to assess a variety of factors about each applying company and build a model that will help predict the chance of success for that venture.

Results

This information comes from the starting ‘charity\_data.csv’ dataset.

* Data Preprocessing
  + Target variable: ‘IS\_SUCCESSFUL’
  + Feature variables: ‘APPLICATION\_TYPE’, ‘AFFILIATION’, ‘CLASSIFICATION’, ‘USE\_CASE’, ‘ORGANIZATION’, ‘STATUS’, ‘INCOME\_AMT’, ‘SPECIAL\_CONSIDERATIONS’, ‘ASK\_AMT’
  + Variables to be removed: ‘EIN’, ‘NAME’
* Compiling, Training and Evaluating the Model
  + Model structure
    - First hidden layer: units = 80, activation = “relu”
    - Second hidden layer: units = 30, activation = “relu”
    - Output layer: units = 1, activation = “sigmoid”
  + Performance
    - The original data construction achieved a starting accuracy of 0.7277
    - Unfortunately, after 3 attempts at optimizing the model, we were not able to achieve the target model performance of 75%.
    - For optimization attempt #1, the columns ‘STATUS’ and ‘SPECIAL\_CONSIDERATIONS’ were removed since they were Bernoulli values and otherwise provided minimal value to the model. This allowed us to increase the accuracy value to 0.7303.
    - For optimization attempt #2, we took changes for attempt #1 and also adjusted the binning values in the ‘APPLICATION\_TYPE’ and ‘CLASSIFICATION’ columns so less values would fall into the ‘Other’ categories. This model resulted in a slightly lower accuracy value of 0.7296.
    - For optimization attempt #3, we returned to the setup of attempt #1 (since it was the most successful thus far) and increased the amount of Epochs used for training the model from 100 to 200. This actually decreased the accuracy of the model down to 0.7296.

Summary

Overall, the models in the various scenarios above were able to correctly predict the funding success of the application ~73% of the time. The majority of the data used as part of this dataset is very vague in nature and/or questionable as to how it would directly impact the success of the funding. These variables include ‘APPLICATION\_TYPE’, ‘AFFILIATION’, ‘CLASSIFICATION’, ‘ORGANIZATION’ and ‘SPECIAL\_CONSAIDERATIONS’. We don’t know much about the specific binned values, so it is hard to use them as a predictive indicator for our output. In order to build a better model, it would help to have more direct performance or statistical information on each company, such as company size, an actual income amount value as opposed to a binned value or a profit margin value. More numerical representations would help to add tangible variables to the model as opposed to the subjective values like ‘APPLICATION\_TYPE’ or ‘CLASSIFICATION’.