SMU Bootcamp – Module 21 Write-Up Report

Overview:

We were tasked with generating a tool for the nonprofit foundation Alphabet Soup that could help select applicants for funding with the best chance of success in their ventures. The original CSV held a dataset that contains more than 34,000 organizations that have received funding from Alphabet Soup over the years. We dropped the non-beneficial ID columns, 'EIN' and 'NAME'. Our target was the column 'IS_SUCCESSFUL' to evaluate whether the money would be used effectively. My final Neural Network has 3 layers containing 15, 10, and 5 nodes, respectively. However, given the specific parameters of this assignment, I was not able to achieve the target of 75% accuracy.

Using the input given by our Professor, Alexander Booth, we were advised to keep the 'NAME' column as it played a factor in the overall success rate. This was examined in my second Jupyter Notebook labeled "Starter_Code-Ryan_Optimized".

Results:

Data Processing –

- What variable(s) are the target(s) for your model?
  The target for the model is the 'IS_SUCCESSFUL' column. It signifies if the money was used effectively.
- What variable(s) are the features of your model?
  The features of this model are the NAME, APPLICATION, TYPE, AFFILIATION, CLASSIFICATION, USE_CASE, ORGANIZATION, INCOME_AMT, SPECIAL_CONSIDERATIONS, STATUS, and ASK_AMT
- What variable(s) should be removed from the input data because they are neither targets nor features?
  The features that could be dropped would be 'SPECIAL_CONSIDERATIONS' and 'STATUS'. There is only a small percentage of cases that had any special consideration, and special considerations cannot be quantified. And 'STATUS' because all rows had the same value of 1.

Compiling, Training, and Evaluating the Model –

- How many neurons, layers, and activation functions did you select for your neural network model, and why?
  In my final model, there are three hidden layers each with many neurons, because this seemed to increase the accuracy closest to 75% (but ultimately not reaching it). The number of epochs changed to 20. The first activation function was 'relu' but the 2nd and 3rd were 'tanh' and the output function was 'sigmoid'. Changing the 2nd and 3rd activation functions to 'sigmoid' also helped boost the accuracy.
- Were you able to achieve the target model's performance?
  No, not with my model and following the module's described directions. However, working with the professor's recommendations, I was able to achieve an accuracy of about 78%.
- What steps did you take in your attempts to increase model performance?
  It required keeping and converting the 'NAME' column into data points, which has the biggest impact on improving efficiency. It also required adding a third layer and using the "sigmoid" activation function for the 2nd and 3rd layers.

SMU Bootcamp – Module 21 Write-Up Report

Summary:

Overall, by increasing the accuracy above 75% we can correctly classify each of the points in the test data 75% of the time. An applicant has an 80% chance of being successful if they have the following:

- The 'NAME' of the applicant appears more than 5 times (they have applied more than 5 times)
- The 'APPLICATION_TYPE' is one of the following: T3, T4, T5, T6, T7, T8, T10, and T19
- The application has the following CLASSIFICATION of: C1000, C2000, C3000, C1200, and C2100.

A good model to recommend is the Random Forest model because Random Forest is good for classification problems. Using this model produces 78% accuracy.