Deep Learning Charity Funding Predictor Report

Christine Li
hli94016@usc.edu

Abstract

1 Overview

Given the background of the non-profit foundation Alphabet Soup wants to create an algorithm to predict whether or not applicants for funding will be successful, we have received a CSV containing more than 34,000 organization that have received funding from Alphabet Soup over the years. With such charity funding information, we would like to compile, train and evaluate a model that can utilize the information inside the CSV file and predict if the applicant of funding will be successful.

2 Data pre-processing

2.1 What variable(s) are considered the target(s) for your model?

Answer: The variable **IS_SUCCESSFUL** is consider the target for my model, which represents whether the applicant will be successful if funded by Alphabet Soup

2.2 What variable(s) are considered to be the features for your model?

Answer: Following variables are considered to be the features for my model.

- NAME
- APPLICATION_TYPE
- AFFILIATION
- CLASSIFICATION
- USE_CASE
- ORGANIZATION
- INCOME_AMT
- ASK_AMT

2.3 What variable(s) are neither targets nor features, and should be removed from the input data?

Answer: EIN, STATUS, and SPECIAL_CONSIDERATION should be removed from the input data.

1. EIN: The value of EIN does not contain any meaningful information to the target, so it can be dropped.

2. STATUS: There are 34294 records with value = 1 and 5 records with value = 0, so it means that most of the records are with status value = 1. and this variable can be removed.

3. SPECIAL_CONSIDERATIONS: There are 34272 records with value = N and 27 records with value = Y, so it means that most of the records are with status value = Y. and this variable can be removed.

3 Compiling, Training and Evaluating the Model

3.1 How many neurons, layers, and activation functions did you select for your neural network model, and why?

Answer: I have selected 4 hidden layers by which the first layer uses 'relu' activation function and the rest three layer uses 'sigmod' activation function. By doing so, I am able to increase the model accuracy to over 79%

3.2 Were you able to achieve the target model performance?

The target model performance is 75%, and yes, the model we built has the accuracy 79.14%, which is greater than 75%.

3.3 What steps did you take to try and increase model performance?

With the baseline starter code given by the instructor, the accuracy of the model is 71.45%, which is far below the target model performance. Here are some steps we have taken to increase the model performance.

1. Drop the EIN, STATUS, and SPECIAL_CONSIDERATIONS columns from the dataframe.

2. Choose a cutoff value = 5 and replace **NAME** whose frequency is less than 5 to other.

3. Choose a cutoff value = 700 and replace **CLASSIFICATION** whose frequency is less than 700 to other

4. Use deep neural net by adding 4 hidden layers, by which the first layer uses 'relu' as activation function, and rest three layer use 'sigmoid' activation function.

In the end, the model we built with the best performance of achieving 79.14% accuracy and 0.460 Loss.

4 Summary

In summary, we have build a binary classification model by firstly dropping some unnecessary columns, replace NAME to other whose frequency is less than 5 and replace CLASSIFICATION to other whose frequency is less than 700, and use deep neural net with 4 hidden layers to achieve the accuracy of 79.14%.

Since it is a binary classification problem, I have also tried other models like Logistic Regression, and Random Forest.

The best accuracy of Logistic Regression model we can achieve is 79.102%, which is greater than the target model performance.

The best accuracy of Random Forest model we can achieve is 77.9825%, which is greater than the target model performance.

Overall, compare with the performance of all these models, the first model I built using deep neural net with 4 hidden layers has the best prediction accuracy and I would recommend using this model.