Report

1. Overview

The purpose of this project is to develop a Neural Network model that may support the Alphabet Soup (A nonprofit foundation) to evaluate an applicant's request for funding and have an initial idea whether the project (from the applicant) would be successful. To achieve that, Alphabet Soup has compiled a dataset with over 34000 entries of successful/unsuccessful projects they have already funded. For each project the following are included.

1. EIN: identification number of the project
2. NAME: name of the organization requesting funding,
3. APPLICATION_TYPE: assigned type of application by Alphabet Soup (T1, T2….etc)
4. AFFILIATION: sector of industry (Independent, Company Sponsored….etc)
5. CLASSIFICATION: company classification (C100, C200, C300…etc)
6. USE_CASE: what the funding is for (ProductDev, Preservation…etc)
7. ORGANIZATION: type of Company (Association, Trust….etc)
8. STATUS: status of project (Active, not active)
9. INCOME_AMT: classification company income (0, 1-9999, 10000-24999…etc)
10. SPECIAL_CONSIDERATIONS: special considerations for application
11. ASK_AMT: amount of money requested.
12. IS_SUCCESSFUL: if the money used effectively and project is successful.

2. Results

2.1 Data Preprocessing

In the current analysis the first two items (columns) in the list provided above are removed (EIN and NAME). These columns are just used as identifiers for each project. In some settings, however, like research grants in Academia, the name of organization/university may play a role in the decision making process based on the organization reputation as well as the outcomes of the past projects.

The rest of the columns in the list apart from IS_SUCCESSFUL are determined to be the model features and they are used to train the model. Column number 12 (IS_SUCCESSFUL) is the target (what the network is trained to predict).

Some features are representative as categorical. Different unique entries for each column are checked, suitable cut-off values are determined. Based on cut-off value for each feature, classifications (0, 1) are established and new columns are added.

In summary

- What variable(s) are the target(s) for your model?
  - IS_SUCCESSFUL
- What variable(s) are the features for your model?
  - APPLICATION_TYPE, AFFILIATION, CLASSIFICATION, USE_CASE, ORGANIZATION, STATUS, INCOME_AMT, SPECIAL_CONSIDERATIONS and ASK_AMT

- What variable(s) should be removed from the input data because they are neither targets nor features?
    - EIN and NAME

2.2 Compiling, Training and Evaluating the Model

Once the Data Preprocessing is completed, the dataset is split into Training and Testing sub-datasets. Entry features for the Training and Testing data set are transformed. About 75% of the original dataset is used to train the model. Summary for training and evaluating the model may be considered by addressing the following questions.

- How many neurons, layers and activation functions are used for the neural network model, and why?
    - A number of models are attempted with different number of hidden neurons, hidden layers and number of training epochs in order to find the best model. The final model contains three hidden layers; 90 neurons in the first hidden layer, 70 neurons in the second hidden layer and 50 hidden neurons in the last hidden layer. Activation functions are set to “selu” in all hidden layers, and “sigmoid” in the output layer. Number of epochs are 300
- Were you able to achieve the target model performance?
    - Target model performance of 75% or higher is not achieved. As per instruction, three attempts are modelled.
- What steps did you take in your attempts to increase model performance?
    - Increasing number of hidden neurons, increasing number of hidden layers, increasing number of training epochs.

3. Summary and Recommendations

In summary the best model shows accuracy of 0.743 and loss of 0.531 on training dataset and 0.726 and 0.580 respectively on the testing dataset. The structure of the model is as indicated in the answer of the first question above.

Looking at the given dataset and determining unique values for each feature, I think features like STATUS and SPECIAL_CONSIDERATIONS should not be included as training parameters. In STATUS only 0.014% of the dataset is shown as 0. In SPECIAL_CONSIDERATIONS, only 0.07% is given as Y. I don’t think these two parameters would add value in the training process. Also, as indicated the Data Preprocessing section, including the NAME of the organization may add value.

To further improve the model, the above may be tried together with, may be, changing and use different activation functions. Also, reducing the numbers of classifications for some of the features may help improving the model performance.

Aside from Neural Networks, and among others, supervised techniques like k-nearest neighbor, random forest, Decision Tree or Extra trees models may, also, be tried