Report: Neural Network Model

Overview of the analysis

The objective of this project was to employ deep learning techniques, such as neural networks, to forecast the success rate of organisations financed by the Alphabet Soup Charity. Alphabet Soup Charity operates as a charitable entity that finances philanthropic ventures globally. The non-profit receives contributions from diverse origins and subsequently dispenses the funds to other charitable institutions. To guarantee optimal utilization of the funds, Alphabet Soup Charity seeks to establish a predictive model that can forecast an organization's likelihood of success based on specific features.

Alphabet Soup's business team has provided a CSV containing more than 34,000 organisations that have received funding from Alphabet Soup over the years. This dataset also contains information about the metadata of each organisation such as affiliated sector of industry and use case for funding. Out of the 34,299 organisations that were funded by Alphabet Soup, 18,261 were regarded as having used the money that they received effectively.

Utilising technologies such as TensorFlow, Keras, and Scikit-learn libraries, a binary classification model was created to predict whether an organisation was successful in effectively using the funds it had received. This report elaborates on the steps taken to build such a model: pre-processing of the data (e.g., addressing missing values; scaling relevant features), the design of its architecture and what was done to optimise the model. To conclude, an evaluation of the model's performance in the form of accuracy and loss metrics is present. Several recommendations for improving the current model are also discussed.

Results

Data pre-processing

Our model's target variable is the IS_SUCCESSFUL column, representing an organization's success status. All other columns, except for EIN and NAME, which serve as identification columns and hold no significance in our analysis, form the model's features. There were 9 such features: 'APPLICATION_TYPE', 'AFFILIATION', 'CLASSIFICATION', 'USE_CASE', 'ORGANIZATION', 'STATUS', 'INCOME_AMT', 'SPECIAL_CONSIDERATIONS', 'ASK_AMT'.

To prepare the data for the deep learning, the following steps were taken:

- Checked for missing values and duplicates within dataset.
- Dropped the EIN and NAME columns as they did not provide any useful information for the model.
- Consolidated low-frequency values in the APPLICATION_TYPE and CLASSIFICATION columns into an Other category. This form of binning helped to reduce the number of unique categories in both of these columns. The purpose of this measure was to prevent overfitting of the model and to increase its generalisability with other datasets.
- Converted categorical data to numeric through the use of the pd.get_dummies function (One-hot encoding). The AFFLIATION, CLASSIFICATION, USE_CASE, ORGANIZATION, and APPLICATION_TYPE columns were converted with this method.
- Before training the model, the dataset was first split into training and testing sets and a StandardScaler was used to normalize the data

Compiling, training and evaluating the model

Building the initial model

A binary classification neural network was created to predict whether an organisation had effectively used the money it had received from Alphabet Soup.

- This deep learning model contained 2 hidden layers with 8 and 5 neurons respectively. The relu activation function was used for these hidden layers.
- The output layer comprises 1 neuron using a sigmoid activation function.
- An 'Adam' optimization algorithm, that used both momentum and adaptive learning rates to speed up the convergence of the optimisation process, was used in conjunction with a binary cross-entropy loss function.
- A minimum of 100 epochs
- During training, the ModelCheckpoint callback function was utilised to save the model weights every 5 epochs.

These parameters were chosen in order to create a simple initial model before adding further complexity during optimisation.

On testing, the initial model attained an accuracy of 72.68%. Unfortunately, this did not meet our anticipated model performance of 75%. This model's weights are saved in the file named: 'AlphabetSoupCharity.h5'.

Model optimisation

To improve the performance of the model, the following steps were taken:

- Dropped the STATUS and SPECIAL_CONSIDERATIONS columns as they did not provide any useful information for the model.
- Added 1 more hidden layer to produce a model with 3 hidden layers.
- Increased the number of neutrons for each hidden layer to 64, 332, 16 respectively
- Continued to use relu functions for each of the hidden layers.
- Added dropout layers than can help prevent overfitting and improve the generalisation of the model.
- Increased epochs to 200 to deepen the model's training
- Changed cutoffs used to both APPLICATION_TYPE & CLASSIFICATION to be less conservative (i.e., to have less categories folded in to an 'Other' category).

To enhance the model's accuracy, we experimented with altering the number of neurons and layers, changing the activation functions, and changing. However, these efforts did not yield a considerable improvement in accuracy (Model 2: 72.72% Accuracy).

Summary & Recommendations.

Despite attempts to optimise the model, the deep learning model was unable to achieve target accuracy of greater than 75%. With an accuracy score of close to 73%, the model seems to provide a reasonable estimation of whether a funded organisation had used the funds it received effectively.

It is worth considering that neural networks may not be the most suitable machine learning model for the current problem given that its value is typically found in image and speech recognition tasks.

Given the characteristics of the dataset (i.e., its large size and classification problem), a different machine learning model such as a Random Forest Classifier or a logistic regression might perform better at predicting which organisation was successful.

If interpretability is a priority, which is often the case when accounting for expenditure by charitable organisations, a logistic regression would be better placed to tackle our classification problem. Considering that we dropped a several uninformative columns in the current attempt, it may be useful to split the entire sample of organisations into sub-sets according to 'ORGANIZATION' type and/or industry ('AFFILIATION') and examine how different features may contribute to the perceived effectiveness of how funds are spent in these organisations. With this reduction in the overall dataset's complexity, a series of logistic regressions can provide a more nuanced picture to inform future funding decisions for organisations in different arenas of industry.