

ANALYSIS REPORT

This project consisted of creating a tool using neural networks and machine learning to help a funder predict which venture to invest in based on their success rates. The data was provided in a .csv file with 11 variables (columns): EIN, NAME, APPLICATION_TYPE, AFFILIATION, CLASSIFICATION, USE_CASE, ORGANIZATION, STATUS, INCOME_AMT, SPECIAL_CONSIDERATIONS, ASK_AMT, and IS_SUCCESSFUL.

11 FEATURES

- EIN and NAME—Identification columns
- APPLICATION_TYPE—Alphabet Soup application type
- AFFILIATION—Affiliated sector of industry
- CLASSIFICATION—Government organisation classification
- USE_CASE—Use case for funding
- ORGANIZATION—Organisation type
- STATUS—Active status
- INCOME_AMT—Income classification
- SPECIAL_CONSIDERATIONS—Special consideration for application
- ASK_AMT—Funding amount requested
- IS_SUCCESSFUL—Was the money used effectively

Data Preprocessing:

What variable(s) are the target(s) for my model?

“IS_SUCCESSFUL” is my target.

What variable(s) are the features for my model?

“APPLICATION_TYPE”, “AFFILIATION”, “CLASSIFICATION”, “USE_CASE”, “ORGANIZATION”, “INCOME_AMT”, “SPECIAL_CONSIDERATIONS”, “ASK_AMT” are the features for my model.

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

EIN and NAME: Drop

STATUS: Filter “STATUS==1” then Drop whole column

Model Compilation, Training, and Evaluation:

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

I defined the model for analysis by determining the number of input features and hidden nodes per layer. I trained the model and tested its fitness through multiple iterations.

Were you able to achieve the target model performance?

Unfortunately, I couldn't optimize my model to achieve the target accuracy. Accuracy : 0.72

What steps did you take in your attempts to increase model performance?

1. Removing T12, T2, T25, T14, T29, T15, T17 in the 'APPLICATION_TYPE' column as the number of each record is too less and it may cause confusion in the model.

2. In the ' CLASSIFICATION' column, dropping where the total record of CLASSIFICATION is less than 20 for the same reason from above.
3. In the ' STATUS' column, filter where 'STATUS'==1 then drop the column as the number of record which 'STATUS'==0 is 5 while 34294 records of 'STATUS'==0

Summary:

The project's preferred accuracy score of 75% was not attained; however, a score of 72% was attained after multiple iterations of the model. A different classification model could be used for future attempts at increasing the accuracy score.

In general, compared to the result in the first classifier model, the optimized model has almost the same rate.

When I designed the first model, I was trying to simplify the data and apply a shallow deep learning model. For each non-numeric-data column, I categorized them into two categories, a category which has the highest count numbers and the other. The result was the first picture above. This dataset is the records about an organization following their existing procedures and established requirements to assess the applicants who can get support or not. However, the deep-learning model is based on supervised machine learning and normally using unstructured data. For example, to classify cats and dogs.