|  |  |  |  |
| --- | --- | --- | --- |
|  | Jeff Pinegar | | jeffpinegar1@gmail.com |
| Assignment: | Deep learning Challenge | 717-982-0516 |
| Due Date: | March 6, 2023 |  |

**Deep Learning Trials**

# Overview

Investigation of

The non-profit foundation Alphabet Soup wants a tool that can help it select the applicants for funding with the best chance of success in their ventures. From Alphabet Soup’s business team, you have received a CSV containing more than 34,000 organizations that have received funding from Alphabet Soup over the years. Within this dataset are several columns that capture metadata about each organization.

* EIN and NAME—Identification columns
* APPLICATION\_TYPE—Alphabet Soup application type
* AFFILIATION—Affiliated sector of industry
* CLASSIFICATION—Government organization classification
* USE\_CASE—Use case for funding
* ORGANIZATION—Organization type
* STATUS—Active status
* INCOME\_AMT—Income classification
* SPECIAL\_CONSIDERATIONS—Special considerations for application
* ASK\_AMT—Funding amount requested
* IS\_SUCCESSFUL—Was the money used effectively

## Objective

Build a model with at least 75% accuracy for predicting success.

## Results

Despite using three different data sets, two different auto-tuning methods, more than 2,000 model fits, 2 to 6 dense layers, and multiple activation functions, I could not achieve 75%. My best effort was 74.7%. The details for this model are covered in “Model Results” below. In very few cases did the accuracy fall below 71%. The results were “trapped” in a very narrow and surprising range.

Below are the details of my investigation.

Contents

[Overview 1](#_Toc128996390)

[Objective 1](#_Toc128996391)

[Results 1](#_Toc128996392)

[Figures 2](#_Toc128996393)

[Tables 2](#_Toc128996394)

[Methodology 3](#_Toc128996395)

[Data Set Preparations 4](#_Toc128996396)

[ASK\_AMT Scaling 5](#_Toc128996397)

[Search Methods and Number of Layers 6](#_Toc128996398)

[Activation Functions 6](#_Toc128996399)

[Results 7](#_Toc128996400)

[Full Data Set 7](#_Toc128996401)

[Typical Results 8](#_Toc128996402)

[Program files 9](#_Toc128996403)

[Model Results 9](#_Toc128996404)

[Half Data Set 9](#_Toc128996405)

[Random Forest 10](#_Toc128996406)

## Figures

[Figure 1: ASK\_AMT Log Scale 5](#_Toc128996407)

[Figure 2: Results for Full Data Sets 7](#_Toc128996408)

[Figure 3: Typical Accuracy Plot 8](#_Toc128996409)

[Figure 4: Typical Accuracy Plot 8](#_Toc128996410)

[Figure 5: Results for Half Data Sets 10](#_Toc128996411)

## Tables

[Table 1: Attempted Models 3](#_Toc128996412)

[Table 2: Layers (Full Data Set) 7](#_Toc128996413)

[Table 3: Layers (Full Data Set) 8](#_Toc128996414)

[Table 4: Layers (Half Data Set) 10](#_Toc128996415)

[Table 5: Layers (Half Data Set) 10](#_Toc128996416)

# Methodology

In the search for a model with accuracy above 75% I tried several data preparations, numbers of layers, activation functions, and search methods; Table 1, summaries those attempts.

Table 1: Attempted Models



## Data Set Preparations

### Features and targets

For this opportunity the target (dependent) variable in the data set is IS\_SUCCESSFUL. The other variables in the data set are potential independent variable or features in our model. Through the cleaning and transformation process some of these variables will be dropped we others will be transformed.

### Cleaning and transformation

This describes the clean of the data except for ASK\_AMT. ASK\_AMT is the only numeric independent variable, so it required special treatment discussed separately

* Data Set One (Minimum Cleaning)  
  This cleaning followed exactly the starter code
  + Dropping Records
    - Dropping records with NaN (none were found)
    - After dropping duplicates (none were found)
  + Dropping Variables that did not contribute to the variance
    - EIN
    - Name
    - STATUS
    - SPECIAL\_CONSIDERATIONS
  + Binning
    - APPLICATION\_TYPE to 9 bins
    - CLASSIFICATION to 6 bins
* Data Set Two (Maximum Cleaning)  
  This cleaning started with Data Set One and then added these steps
  + Dropping Variables that did not contribute to the variance
    - STATUS
    - SPECIAL\_CONSIDERATIONS
  + Binning
    - APPLICATION\_TYPE to 6 bins
    - CLASSIFICATION to 6 bins
    - USE\_CASE reduced to 3 bins
    - ORGANIZATION reduced to 3 bins
  + ASK\_AMT
    - Dropped records with ASK\_AMT greater than $1 billion
    - Dropped records with ASK\_AMT more than 50 times the (INCOME\_AMT)
    - Dropped records with ASK\_AMT with INCOME\_AMT = 0 and ASK\_AMT greater $100,000
* Half Data Set (Maximum Cleaning)  
  I took a 50% random sample of Data Set Two. I erroneously split the data set in half early to correct overfit (I know this is the wrong cure for overfit); however, using this data set, I got a model with 75% accuracy (but it was a random chance), that could not be reliably replicated. Therefore, I wanted to do a more robust search.

## ASK\_AMT Scaling

Both data sets still presented ASK\_AMT with a huge range that needed to be managed. I attempted two methods to wrangle this data

* Standardization (Z-Score)
  + After standardization, the Z-Score ranged from 0 to 45!
  + Dropped records with ASK\_AMT Z-Scores greater than 3 (112 records)
* Logarithmic transformation
  + I believe this is the best approach for this data; see **Error! Reference source not found.**. Nevertheless, I did the analysis with both transformation methods.

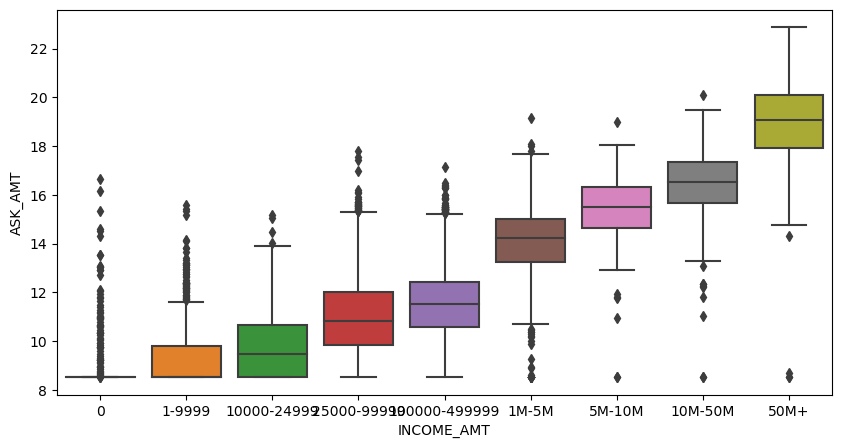


Figure 1: ASK\_AMT Log Scale

## Search Methods and Number of Layers

I used three types of search methods to fit the model.

* Expert Opinion  
  After considering the data set, I tried various combinations of layers, neurons per layer, and activation functions that I thought would work for this data.
  + Number of layers from 2 to 3
  + Number of neurons per layer
    - Layer 1: 1x, 2x, and 3x the number of inputs were evaluated
    - Layer 2: 1/2, 1/4, and 1/8th of layer 1
    - Layer 3: 2, 4, 8, 16
  + Activation functions
    - Output layer was Sigmoid, since we are attempting to classify two states.
    - Interlays we mostly relu, but with more failure, I tried Sigmoid also
* Keras\_tuner
  + Using Google Colab – The problem with Colab is that my session would time out before the model would fully evaluate. The best results, when the various attempts ended were all between 71% and 73.5%. I used two different tuners for this search:
    - Tuner = Hyperband
    - Tuner = RandomSearch
  + Using my PC, I could run longer (more than 8 hours). Unfortunately, these attempts also eventually crashed. The Keras\_tuner writes a file to disk with every pass. Eventually, I would get a write error, and the process would stop. The best results, when the various attempts ended were all between 71% and 73.5%. I used two different tuners for this search:
    - Tuner = Hyperband
    - Tuner = RandomSearch
* Exhaustive
  + To avoid the write file problem I was experiencing with Keras\_tuner I built a nested for-loop model where I could control the number of layers and neurons as well as the activation functions. I tried 4 inner layers with the following numbers of neurons (124, 62, 31), (20, 10, 40), (0, 5, 15), and (0, 6). Each could have an activation function of relu or Sigmoid. This resulted in a total of 864 combinations. I ran them all.

## Activation Functions

* Output layer - In nearly all cases, the output function was sigmoid. However, I also tried tanh, relu, and softmax as output activation functions. Sigmoid and Tanh were selected because of their ability to make the final classification for IS\_SUCCESSFUL or not.
* Dense layers – To begin with, I focused on Relu. However, all the automated testing (Keras and exhaustive use both Relu and sigmoid activation functions in the dense layers.

# Results

## Full Data Set

I recorded 1,118 unique model fittings with the full data set (not the 50% sample). Of the those, 1,030 trials had accuracy scores greater than 0.7, Figure 2 is a histogram of the results (with a bucket for less than 72.23%).

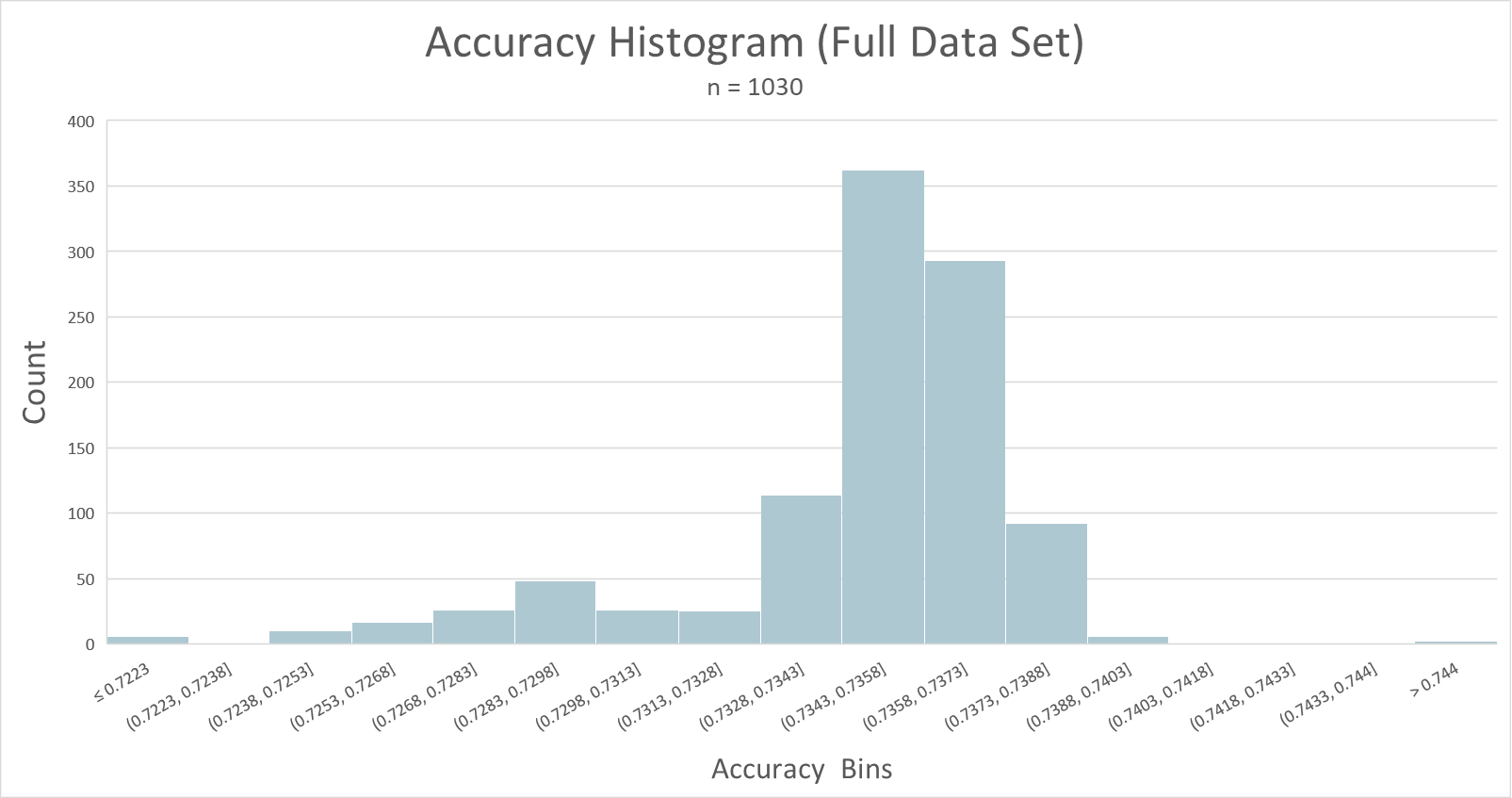


Figure 2: Results for Full Data Sets

The limited range of results is surprising, given the range of models that were evaluated. Table 2 identifies the minimum and maximum number of layers tested, while Table 3 identifies the frequency of activation functions tested. The spreadsheet Report figure.xls as the details for the 1,118 runs.

Table 2: Layers (Full Data Set)



Table 3: Layers (Full Data Set)



## Typical Results

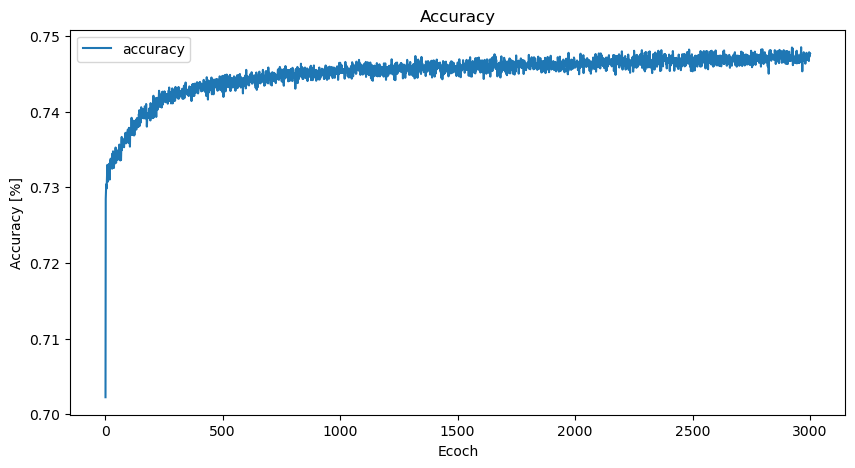


Figure 3: Typical Accuracy Plot

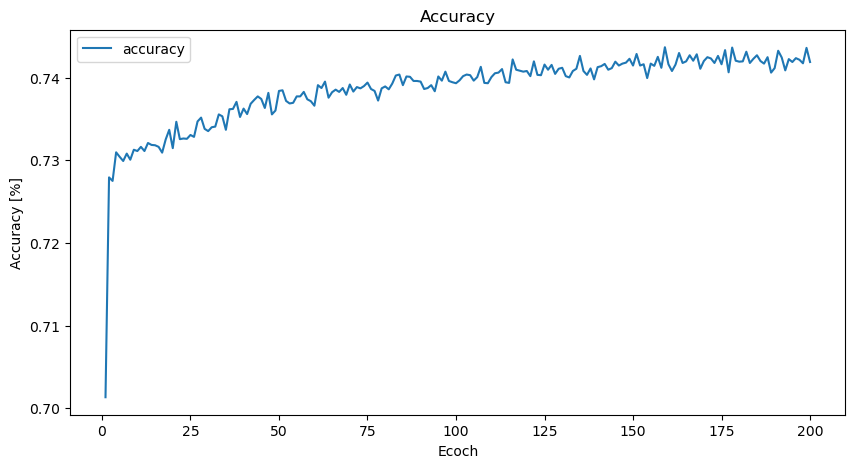


Figure 4: Typical Accuracy Plot

### Program files

These Jupyter notebook files are typical of the code that was created and used for this analysis. I say typical because there were many variations used and discarded because they did not make a substantial improvement over other models. See Figure 5: Accuracy Plot for 74.7% Model and Figure 6: Loss Plot for 74.7% Model

* JSP DeepLearningChallenge\_Min\_Clean.ipynb (Single pass)
* JSP DeepLearningChallenge\_Final\_Solution.ipynb (Nested for loop model)
* JSP\_21\_autotune 003.ipynb (Keras\_Tune for the PC)
* JSP\_21\_colab\_autotune 002.ipynb (Keras\_Tune for Google Colab)

### Model Results

The single pass jupyter notebook was use to build this models after exploratory research was completed using the other program to make selections for the model. I then ran that configuration with 3,000 epochs. The result was a model with 74.7% accuracy.

* 2023-03-05-221906-7234-Min clean 747 model.h5
* 2023-03-05-221906-7234-Min Clean 747 weights.hdf5

My best (most accurate) model had the following parameters

* inputs = 40 Number of independent variables in the data frame
* lay\_1\_n = 102 Number of neurons in the first layer
* lay\_2\_n = 10 Number of neurons in the second layer
* lay\_3\_n = 10 Number of neurons in the second layer
* lay\_4\_n = 2 Number of neurons in the second layer
* act1 = "relu" Layer 1 activation function
* act2 = "sigmoid" Layer 2 activation function
* act3 = "relu" Layer 3 activation function
* act4 = "relu" Layer 4 activation function
* actout = "sigmoid" Output activation function

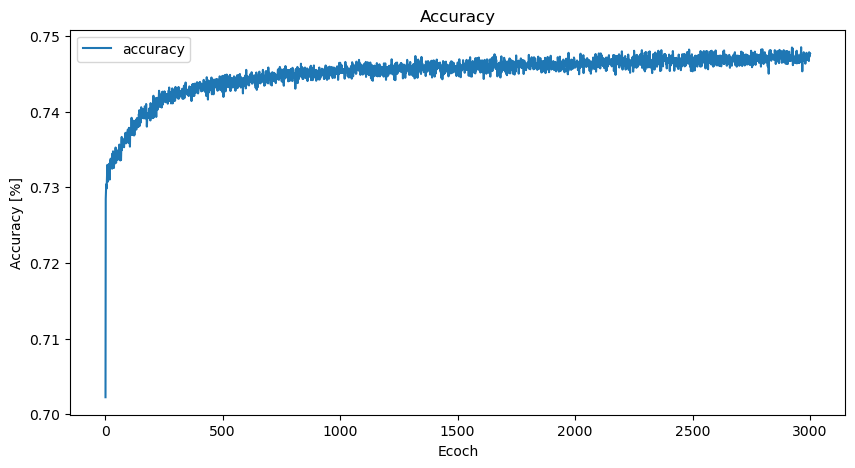


Figure 5: Accuracy Plot for 74.7% Model

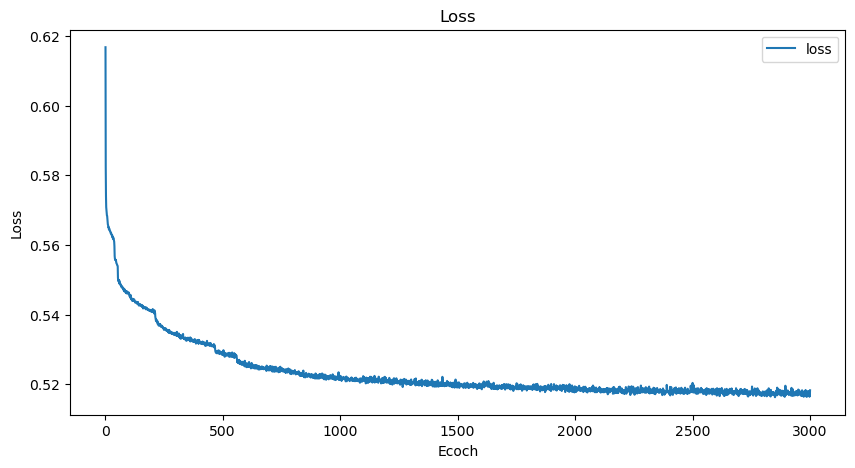


Figure 6: Loss Plot for 74.7% Model

Half Data Set

I recorded 695 unique model fittings with the half data set. Figure 5 is a histogram of the results. The limited range of results is the same as the with the full data set. Table 4 identifies the minimum and max number of layers tested, while Table 5 identifies the frequency of activation functions tested. The difference between the full data set and the half data set is most likely due to the number of epochs. The full data set was mostly run with epoch = 50 while the half data set was run with epoch = 20.

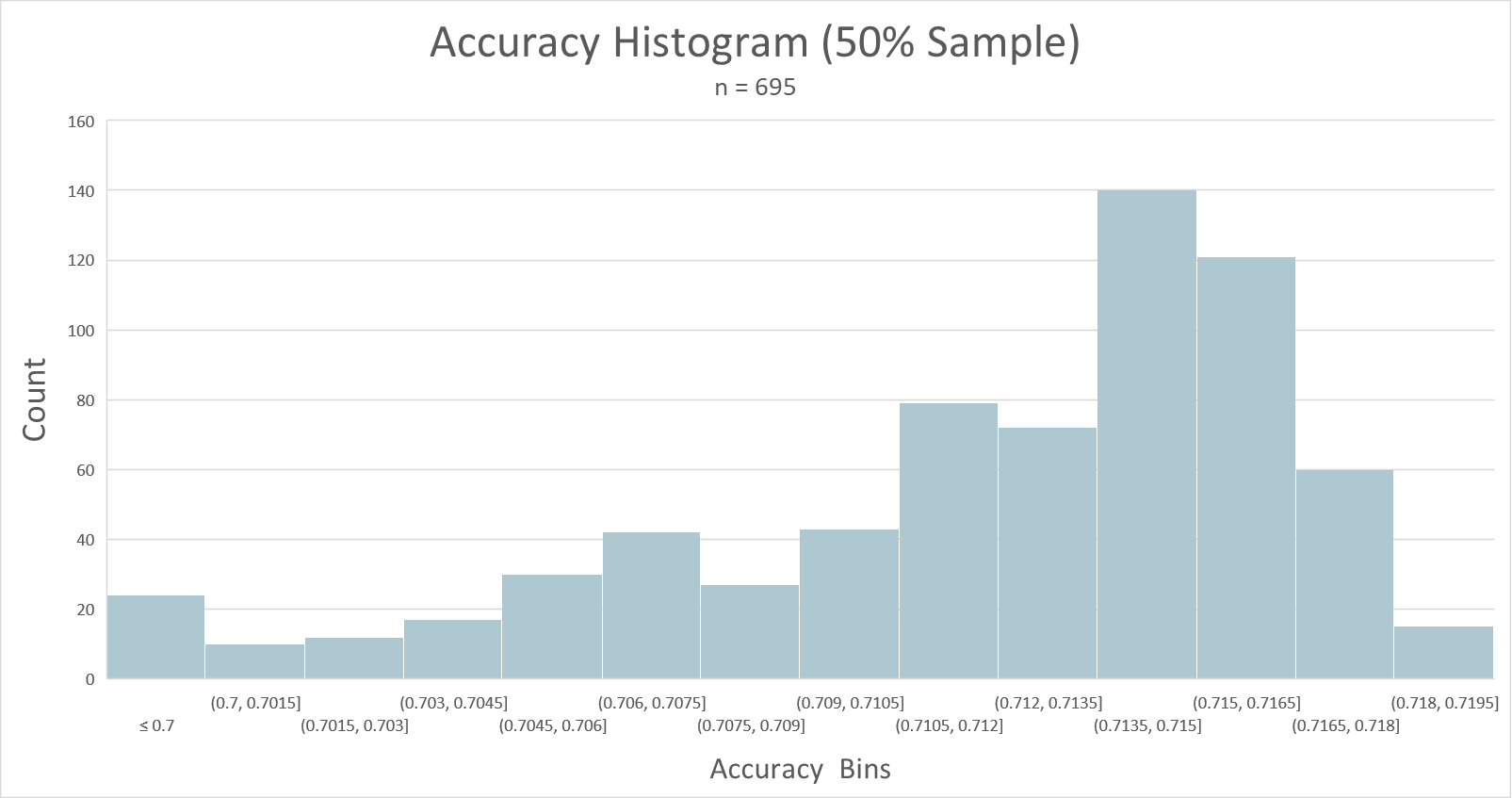


Figure 7: Results for Half Data Sets

Table 4: Layers (Half Data Set)



Table 5: Layers (Half Data Set)



## Random Forest

In addition to deep learning neural networks, I attempted a RandomForce model. The results were no better than the neural network solution.

True positives (TP): 3588

True negatives (TN): 2563

False positives (FP): 1474

False negatives (FN): 950

precision = 0.70881

accuracy = 0.71731

sensitivity = 0.79065

F1 = 0.7475

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | recall | f1-score | support |
| 0 | 0.73 | 0.63 | 0.68 | 4037 |
| 1 | 0.71 | 0.79 | 0.75 | 4538 |
| Accuracy |  |  | 0.72 | 8575 |
| macro avg | 0.72 | 0.71 | 0.71 | 8575 |
| weighted avg | 0.72 | 0.72 | 0.72 | 8575 |