Assignment 3

Topic: Performing ANN on a given dataset

CSCI 4155

By Shadman Mahmood

B00780608

Objective

In this assignment we had to implement Artificial Neural Networks on the dataset provided to us called the lending_club_loan.csv. The goal for our model was to be able to assess if a customer is likely to pay back their loan.

Detailed Process

We were asked to visualize loan_status column so by using seaborn.countplot I was able to analyze it. Next I performed df.corr to show the correlation and fed these values into seaborn.heatmap to visualize the results in the form of a heatmap. No there was no duplicate data, installment showed a high correlation with loan amount but it is different from the actual loan amount so I decided to keep the data. Then to analyze the relationship between loan_status and loan_amount I used seaborn.boxplot to create a boxplot, and I could conclude that fully paid and charged off statuses shared a similar amount range. By calculating the summary statistics for this data I could further concur that their means were very close to each other. By finding the unique grades and subgrades I could see there were 7 categories of grades as well as 4 categories for each grade as the subgrade (i.e. A1, A2, A3, A4 …..G4, G5). After plotting the countplots for the grade and subgrade column we could see the lowest likelihood to pay back loans were in the “F” as well as “G” category.

By checking for null values, we could see emp_title, emp_length, title, revol_util, mort_acc, and pub_rec_bankruptcies had missing values. Since emp_title and title had too many categories to perform encoding, so they were dropped. Since revol_util and pub_rec_bankruptcies had a small amount of missing data so missing data could dropped. Emp_length percentage of people who were charged off was relatively quite close to each other so it could be dropped as well. For mort_acc because of the next question I decided to use correlation function to check for any columns that might correlate to mort_acc. The closest match was by total_acc so the mean of that column was used to impute the missing data. Grade column could also be dropped since we have subgrade column which provides a more descriptive information about grades. Additionally, I dropped issue_date as our model is going to predict if a person will pay their loan before the company gives it out so generally, we won’t have access to the issue date. All non-numeric value columns were converted to dummy variables and joined to the dataframe and the corresponding nonnumeric columns were dropped.

Lastly, I split the training and the test set with the specified ratio of 0.2 for test_size and random_state of 101. After normalzing the data I created a ANN with 1 input layer, 2 hidden layers and 1 output layer using the 78 neurons as we have 78 features. After testing, evaluating and saving the model I ran the test customer information given to us. The resultant output was that the customer is likely to pay their loan amount.