Internet Advertisements

- **Summary:**

The objective is to create a classifier model which can find out whether an image in a web-page is an Advertisement or not based on its features. A research paper 'Learning to remove advertisements' by Nicholas Kushmerick was studied to gain proper insights on the features of data. Noise removal and pre-processing of data was done. After pre-processing steps, a feed forward neural network-multiple layer perceptron was used for the classification model. Total data set was divided into training, validation & test data sets. Finally training and testing of the model was done using respective data sets. The model accuracy was found out to be 98.4 percent for test data.

- **Details:**

1. **Specific Characteristics of Data set:**
   - ➔ The given data set contains 3279 samples each of 1559 attributes. The data set consisted of 459 instances of 'ad' class and 2820 instances of 'nonad' class.
   - ➔ If the aspect ratio is greater than 4.5833, alt doesn't contain 'to' but contains 'click+here', and $\mathbf{U_{dest}}$ doesn't contain 'http+www', then instance is an AD.
   - ➔ If U_{base} doesn't contain 'messier', and U_{dest} contains the 'redirect+cgi', then instance is an AD.
2. **Data Preprocessing:**
   - ➔ Data preprocessing is needed as 28 % of data of 3 parameters are missing in the given data set.
   - ➔ Preprocessing steps include:

     a) Handling missing values.

     b) Converting string target values to numerical to make the data ready to be processed.

     c) The target class values were converted to 0s and 1s to make the data set ready for processing.
3. **Handling Missing Values:**
   - ➔ Missing values were found in column# 1, 2, 3, and 4(namely height, width, aratio, local). Details about the missing values are as follows:

Column Name	# missing values with class 'ad'	# missing values with class 'nonad'	# total missing values
Height	830	73	903
Width	828	73	901
Aspect Ratio	837	73	910
Local	10	5	15

- ➔ Missing height and width cells were replaced with the mean of the available height & width values of all the samples belonging to the same class. For example, if height is missing for a sample belonging to the class “ad”, then the mean of available height values of all the samples belonging to class “ad” is placed in the cell. Similar procedure was followed for width column also.
- ➔ As all the missing height and width cases were solved, the missing Aspect Ratio values were calculated by finding width/height.
- ➔ Missing Local cells are replaced with that value which is in majority in the available aratio values of all the samples belonging to the same class, because Local column can take only values 0 or 1. For example, if Local value is missing for a sample belonging to class “nonad”, then the value which is in majority in all the samples belonging to the class “nonad” is placed in the cell.

4. Prediction Model using Neural Network:

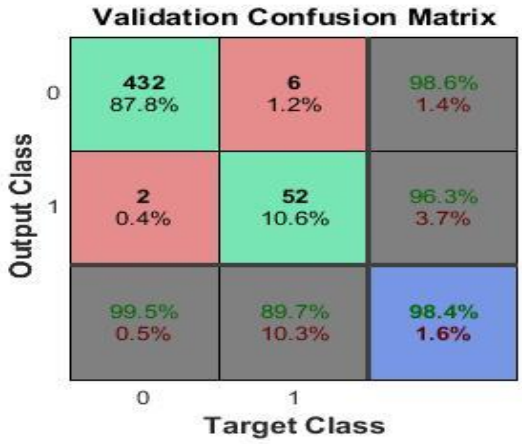
- ➔ A 2 Layer Feed Forward (a hidden layer consisting of 200 perceptrons and an output layer consisting of 1 perceptron) network was used for the model. Sigmoid function was chosen as perceptron excitation function. Error correction method was Conjugate Gradient Descent Back propagation. The total data set was divided into three parts randomly (maintaining the original ratio of 459:2820 for ‘ad’ to ‘nonad’ classes).

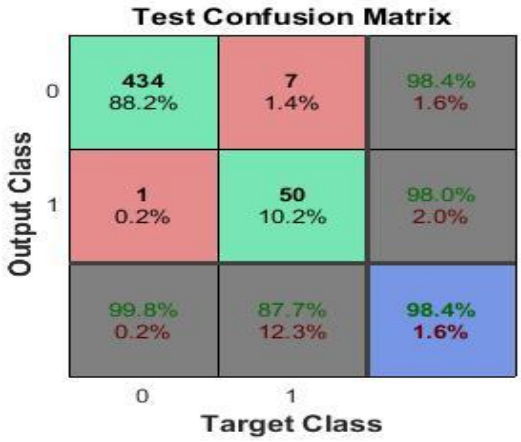
  Train – 70 %
  Validation – 15 %
  Test – 15 %

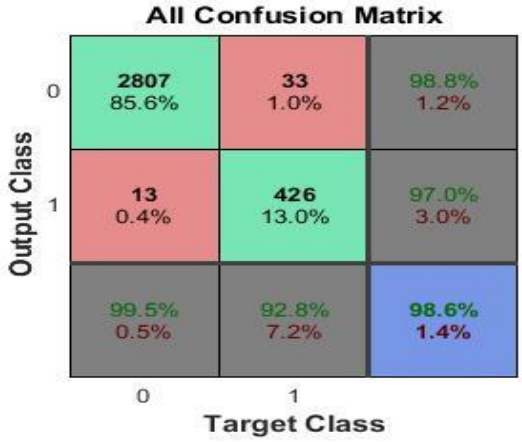
- ➔ Then the network was trained and tested upon the test data.

5. Evaluation of Results:

➔ The Confusion Matrix is as shown below:







➔ The Receiver Operating Characteristics is as shown below: