HARMANTEPE Melis
EZEDINE Marcel
Group A11

Model Analysis

Prediction task is to determine whether a person makes over 50K a year.

1. Hierarchical Clustering

In this model the aim is to identify the clusters that are the closest to each other, then to merge the most similar clusters together until we form one big cluster. This is visible thanks to a dendrogram. Once we have our dendrogram, we find the longest vertical line that is not cut by any other horizontal lines. We cut this vertical line at its upper and lower extremities to define the number of clusters. The number vertical lines that cross this new line will give us the number of clusters. In our case, the number of clusters is 4. From the agglomerative clustering graph we can observe that there are almost the same amount of people around the ages 30-40 that makes either 50k or less than 50k per year. We can vaguely say that almost more people from the age group 55-90 doesn't make 50k per year. For the rest of the age groups as well, this graph is not efficient in terms of analyzing and predicting the data. I can say that this model is the worst of all three.

2. Decision Tree

Our tree learns whether income exceeds $50K per year based on census data. Our dataset includes categorical features as well as numerical values. Our target feature is "income". For this reason, the best "way" to split the dataset such that dataset 1 contains incomes <50k and dataset 2 contains >=50k is that we fix the target value as "income". Once we compute the accuracy by comparing actual test set values and predicted values, we get a classification rate of 82.78%. So, we can optimize this graph and the accuracy by fixing a maximum depth of 3 for the tree. We observe that the tree is way more visible but the accuracy has went down by 1 percent, reaching 81%. Even though the accuracy has decreased, it is not a big drop but the graph is more visible so this is the best, adaptable model to predict and visualize the dataset that we are working with. It not only gives the best accuracy but also offers visibility.

3. Artificial Neural Network

Here, we are writing an Artificial Neural Network to find out whether a person makes over 50K a year. After preparing the dataset, we store the dependent value/predicted value in y, which is the column "income". Since we haven't put any parameter in the Sequential object, we will be defining the Layers manually. So, to define the output layer dimension, we used a rule of thumb which takes the average of the number of nodes/attributes in Input and Output Layer (13 +1)/2 =7). Then we compile and make a prediction. Our prediction on Test set results tell us that the accuracy will be more than 0.5. Then we make our confusion matrix and print out the accuracy which is 75.43% and good enough. So this model by far is not the best to predict because we get a higher accuracy score with the decision tree.