Problem statement: The main aim of this project is to identify customers who are potential for the business deal. I considered **the bank-additional-full.csv** file as the data collected from the past year and the bank-additional.csv as the present year's file. I used the **bank-additional-full.csv** file for training purpose and the **bank-additional.csv** file for test purpose.

Data description and model selection: There are 20 input features and 1 output feature. This is a classification problem statement with 2 possible outputs i.e., either yes or no. Based on my observation on the data I noticed that the data is **Biased** i.e., there are too many 'no' outputs compared to 'yes' outputs. The total number of 'yes' in training data is around 4000 and the total number of 'no' in the training set is around 38000. These figures tell that the data is biased towards 'no'.

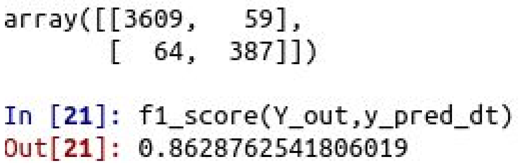
Because of this biased data, I chose the **Decision tree classifier**. The read in a publication that the Decision tree classifier is much better for training and testing in biased data.

Performance:

F1 score is the best measure for performance when we have biased data.

F1 score=(2*precision*recall)/(precision+recall)

The F1-score for my model is 86.8%.
The confusion matrix and F1 score for my model are attached below.



```
array([[3609,   59],
       [  64,  387]])

In [21]: f1_score(Y_out,y_pred_dt)
Out[21]: 0.8628762541806019
```

In testing data
Total cutomers=4120
Total yes's=451
Total no's= 3668

If we consider a cost of 10 units per call and 100 units if the customer says 'yes' then if we call all the people for the test data then the

We earn a total of 45100 units but we spent a total of 41200 units
So we lose a total of 36680 units. (useless calls)

But if you make calls based on my model you will earn a total of
38700 units but you spent a total of 4460 units so the loss is only 590 units(useless calls).

So the profit percentage if you call all the customers is 9.4%. [(45100-41200)/41200]
But the profit percentage for my model is 767%. [(38700-4460)/4460)