Problem Statement:-

The marketing team wants to launch another campaign, and they want to learn from the past one. You, as an analyst, decide to build a supervised model in R and achieve the following goals:

• Reduce the marketing cost by X% and acquire Y% of the prospects (compared to random calling), where X and Y are to be maximized

• Present the financial benefit of this project to the marketing team.

Dataset

The dataset and dataset description can be obtained from here [https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

Solution

Since the target variable ‘y’ is binary, we need to use classification algorithms in the supervised machine learning model. Decision tree is the best algorithm for this case since the dataset is such that simple splits can lead to an accurate result. This observation is obtained from the plots drawn (present in the code) such as duration vs y, age vs y, nr.employed vs y, etc.

Now by simple conversion of factors to numeric values and implementation of decision tree as shown in the code, we obtain an accuracy of 92.6%. But, if we plot the decision tree we see that ‘duration’ is very important factor used in splitting, **but since the marketing team wants to set up a new marketing campaign and needs to predict customer responses prior to making calls, then ‘duration’ factor cannot be used here. Since the duration of the call can only be obtained after making the calls therefore such factors are irrelevant to prediction.**

Hence, ‘duration’ column is first removed from the data set. Now, after applying the algorithm as shown in the code, we get **an accuracy of 90.56%**. Although it is less than the previous case but since the factors present in the model are all relevant in this case, this model makes more sense than the previous one and will surely perform better in realistic predictions.

Now coming back to the problem statement, in case of random calling, i.e. if we call all our customers irrespective of the prediction, we would have to make 41188 calls,

Assuming that every call requires ‘a’ unit of marketing cost (telephone bill,calling executive salary,etc)

Total Marketing cost (random calling) = 41188a

And if we use the model for predictions, we can see from the confusion matrix:

pred no yes

no 0.82180141 0.02767662

yes 0.01189609 0.08315125

The percentage of correct prediction is (0.82+0.09 = 0.91) 91%.

Now the marketing team will call only the customers for which our model predicts ‘yes’.

Marketing cost(in this case) = (0.119+0.083)\*(41188)\*a = 8320a

So if we go according to our predictions, we can reduce the marketing cost by x percentage:

X = (41188-8320)/41188 = 0.79 or 79%.

And we will successfully acquire y% of prospects:

Y = (0.083)/(0.083+0.028) = 0.75 or 75%.

**Model can be further modified so as to correctly predict ‘yes’ so that more prospects can be acquired with a little less reduction in marketing cost.**

As it is clear from the above evalution, using model reduces 79% of our marketing cost and we cover 75% of the prospects. Therefore, it is very beneficial financially.