LEAD SCORE CASE STUDY SUMMARY

PROBLEM STATEMENT-

We are given following problem statement-

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

This is classification machine learning problem. In this case study, we must determine **whether the given lead is highly likely to convert or not** based on given information.

DATA EXPLORATION-

First, we did **Exploratory data analysis**, in which we have found out that there are certain features which have **high number of nan values** and some have **low number of nan values**. We tried **to impute values with 'missing' and -1** which have a high number of missing values. We did this so that we do not lose our data.

Other features which have a smaller number of nan values, we **imputed with median values and most frequent value**. Then, we did scaling of numerical features and one hot encoded the categorical features.

MODELLING-

In modelling part, we used **logistic regression**, subsequently we also used **k-fold cross-validation** technique just to get an idea how our model is performing. In results we came to know our model is was **generalizing well**.

EVALUATION-

Next, we calculated **confusion matrix, precision and recall scores**.

In case study, it was mentioned the company wishes to identify the most potential leads, in order to do that we might want to have **high precision** so that our model gives high potential leads only.

In order to do that, we **increased the threshold** of our model and finally we get **decreased false positives** in our model.

Finally, we also evaluated our model on test data in which we get around **92% accuracy**.