LEAD SCORE ASSIGNMENT

- *SUMMARY REPORT*

In the given assignment, we have cleaned and prepared the data. Checked and worked on any outliers. After this point, the data was cleaned, and around 98.5% of the data was retained. After this univariate analysis was carried out, some of the following insights emerged:
1. There has been an overall conversion rate of around 38%.
2. The majority of leads originate from landing page submissions. Maximum conversion also happened from landing page submission, with a conversion rate of around 36.20%. Lead Add Form also originates total 608 with a conversion rate of around 98.60%.
3. Google is the source that gives the most leads, with a conversion rate of around 40.42%. Leads that are coming from the Welingak website source have the highest rate of conversion, around 98.4%. Leads that are coming from the reference source also have a huge conversion rate of around 92%.
4. Most of the conversion has happened through the emails that have been sent.
5. Most conversions happened when calls were made. However, it can also be seen that two leads opted for "Do Not Call", but they still got converted.
6. The last activity value of 'SMS Sent' had maximum conversion with a conversion rate of around 63%.
7. In "What is your current occupation?" we can see that maximum coverage takes place among the section of people who are unemployed. Working professionals have a huge conversion rate. In total, nine out of nine housewives converted. Out of 8 businessmen, 5 got converted, and 9 out of 15 got converted from the other section.

After this, we converted binary variables (yes/no) to 0/1 and then created dummy variables and dropped repeated columns. Resulting in the dataset having numerical values. From here, we started the modelling phase, where we tested and trained the data. This led to a model with an accuracy rate of around 81%, which is good. However, we will also need to calculate the other metrics, as we cannot depend only on the accuracy metrics. Such as checking VIFs and calculating metrics like sensitivity, specificity, false positive rate, positive predictive value, and negative predictive value.

We also created the ROC curve. An ROC curve demonstrates several things:
1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
4. 0.37 was the optimum point to take it as a cutoff probability.

The conclusion after training and testing was that, while we have checked both sensitivity-specificity as well as precision and recall metrics, we have considered the optimal cutoff based on sensitivity and specificity for calculating the final prediction. The accuracy, sensitivity, and specificity values of the test set are around 79%, 77%, and 80%, which are approximately closer to the respective values calculated using the trained set. Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80% Hence, overall, this model seems to be good.