CASE STUDY SUMMARY

Approach:

Framing the Problem: We want to solve the business problem of identifying potential leads and increase the rate of lead conversion by prioritising those leads in an efficient way.

So we have two specific objectives here:

1. Building a binary classifier model which classifies new customers into potential 'Hot leads' (likely to buy) or 'cold leads' (not likely to buy).
2. Highlight features which are important for the classification so as to provide business relevant solution and actionable insights.

Analysing The problem: While some of the features in the dataset are directly input by customers (such as specialisation, current occupation etc) , some are concerned with customer activity(Total visits, total time spent on website etc) tracked by X education. Apart from these, some features captured the perception of X education about the potential leads (Lead quality, asymmetric profile/activity scores etc), which are bound to be biased and hence we dopped such features from our consideration.

After **checking duplicate records** (no duplicates found) and **missing value treatment**, we performed **EDA** to gain more understanding about the features. we removed some more features which weren't adding value to our analysis.

Afterwards, **converting the categorical features into numerical** ones, we have **built several iterative models** in train set such that multicollinearity remains low and all the selected features are statistically significant. Identifying the important features, we proceeded to test our logistic regression model on unseen test data.

Analysing The Solution: In the first logistic regression model, we assumed the threshold probability value as 0.5 based on which we have got accuracy, sensitivity, specificity as around 81%. But as the problem suggests, we are more concerned **with identifying potential hot leads and increasing lead conversion rate**, it is important to base our evaluation on the metrics which evaluate the positive prediction rate such as **precision and recall**. So, we have calculated an optimum cut off value by **plotting precision recall trade off curve**, based on which both the metrics have come out to be around 77% for unseen test data, which aligns with our client's expectations.

Observations & Suggestions:

While preparing the data and building the model, we have come across following observations:

1. **Total time spent on website** is very influential attribute for hot leads.
2. When **lead origin** is thru **lead add form / landing page submission**, leads are likely to be converted.
3. When **customer source** is **olark chat or welingak website**, they are more likely to buy courses.
4. Customers having **hospitality management as specialisation** are more likely to enrol.
5. **Working professionals** have more inclination to upskill themselves thru such courses than other categories.
6. When customers are involved in **olark chat conversation or sms conversations** as **last notable activity**, they found to be more likely to buy courses from X education.

Keeping the above in consideration, X education can focus on **potential hot leads thru efficient channels** and assistance to ensure leads get converted and thus result in **increase in lead conversion rate.**