Summary:

This analysis is to discover approaches to identify the potential buyers and focus more on them to opt for their courses from the given data. The data consists of customer visit their website, the time they spend and the details they file and add they watch etc. The below are the approaches and steps performed to achieve the final solution.

1. **Data Cleaning:**

The value 'select' present in the place of missing value, hence replace them with null value. Checking for the univariate columns to drop as these are not mandatorily be present for Analysis. Most of the column in the data frame has more than 35% null values, hence dropping them except specialization which we wanted to infer some insights. Where null values are replaced with "Not Provided".

When checking the value counts for country column not many from outside, hence the elements were changed to 'Outside India', and 'Not Provided' for null values.

2. **EDA:**

Most of the columns are categorical and an EDA has been done to check the state of our data and based on our analysis these categorical values are not much important however few points are inferred from these analysis

i) Working professionals are potential buyers as they opt for better career progress and unemployed also turned to be a potential buyer.

ii) Leads are mostly getting converted through 'SMS sent' as we could observe this in EDA.

The numeric values appear to be acceptable and no abnormalities were observed.

3. Dummy Variables:

The dummy variables were created and later the original columns are removed also 'not provided' elements were eliminated. For scaling the numeric values, we used Min-Max Scaler.

4. Train-Test split:

The Train and Test set split was done at 70 – 30 percentage from the data frame.

5. Model Building:

RFE was done to achieve the best 15 variables and the rest of the variables were eliminated manually depending upon the P –values and VIF values. We ensure at the end of model building the VIF is less than 2.5 and P value is less than 0.05.

6. Model Evaluation:

We made a confusion matrix from the final model and use the cut-off 0.35 as it gives the accuracy 81%, sensitivity 70%, and specificity which came to be around 88%.

7. Prediction:

The prediction was done on the test data frame and with an optimum cut-off of 0.35 with accuracy 81%, sensitivity 81%, and specificity of 80%.

8. Precision-Recall:

Again with the same optimal cut-off of 0.35 we could achieve 79% precision and 70% recall. Which is pretty good.

Also with precision-recall curve, the optimal cut-off is 0.41, we have accuracy of 81%, sensitivity/recall is 75%, and specificity is 78% and precision 71%.

9. Conclusion:

By utilizing the 'conversion_Prob' values we found Lead score for each variable also the features are identified which contribute most to a Lead getting converted successfully.

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

The variables given below in descending:

1. Last activity is Olark Chat Conversation
2. Current occupation_student
3. Lead Origin_lead add form
4. Lead sources are
    i). Olark Chat
    ii). Welingak Website
5. Total time spent on website is more
6. Total Visit on website is more
7. When the last activity is SMS sent and Do not Email is Yes
8. When their current occupations are Unemployed and a working professional.

Keeping these variables in mind the X Education can develop their business by focusing these potential buyers to buy their courses.