Summary

This analysis is done for X Education in order to determine how to attract more business professionals to their courses. We learned a lot from the dataset provided on how potential customers visit the site, how long they stay there, how they get there, and their conversion rate.

The following technical steps were used:

1. **Reading & Understanding the data:**
    - Importing necessary libraries
    - We observe the variables.

2. **Handling missing values:**
    - Removing the 'Select' variable as it has no meaning in the data.
    - Dropped the high percentage of null values more than 40%.
    - Checked for number of unique categories for all categorical columns.
    - Identified and dropped highly skewed columns.

3. **Treating outliers & Imbalance data**
    - Treated missing values.
    - Detected outliers.
    - Exploratory Data Analysis was done to check the condition of data.

4. **Visualising the data**
    - Plotted a heat-map to look at the correlation of numerical variables.
    - We found that most of the leads don't want to be emailed about the course.
    - We also found that Non-conversion rate is more than conversion rate in every specialisation.
    - Most leads are from Mumbai.

5. **Data preparation**
    - Here we do the mapping and bring in the dummy variable.
    - Later we get the shape of the data-set after adding dummies and dropping few columns.

6. **Splitting into Train-Test Set**
    - We split the data-set into 70-30 ratio.

7. **Rescaling**
    - We find out the lead conversion rate which is approximate 38%.

8. **Building a Linear Model**
    - First we build the model with all the columns.
    - Later by using RFE with 15 variables.
    - Irrelevant features were removed manually depending the VIF and p values (VIF < 5 and p-value 0.05).

9. **Plotting ROC Curve**
    - A confusion matrix was made, later using the ROC curve or area and was used to find the accuracy, sensitivity and specificity which came around 0.93 which is ~1 which means that the model is good.

10. **Finding Optimal Cutoff Point**
    - Here we obtain the balanced sensitivity and specificity.
    - Here we build confusion matrix and find out accuracy score which comes ~0.83.

11. **Precision & Recall**
    - Precision = TP/(TP+FP)
    - Recall = TP/(TP+FN)
    - This model was used to recheck and to find the cut-off point of 0.42.

12. **Prediction & Evaluation on Test-Set**
    - Hence we can see that the final prediction of conversion has a target rate of 86%, which is more than expected (80%).
    - From all this we can say that this model seems to predict the conversion rate very well and we should be able to give this model to the CEO for making good calls.

13. **Conclusion**
    - The variables that matters most are:
        1) Total time they spent on website
        2) Total number of visits
        3) When the lead source was: Google, Direct Traffic, Reference, Welingak website
    - Whether their current occupation is working professional or if they are unemployed.
    - When the lead origin is lead add form.

- When was the last activity:
  1) SMS sent
  2) Email opened.
- When tags was:
  1) Closed by Horizon
  2) Lost to EINS