An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

The task is to create a lead scoring algorithm. The following steps are used for this task:

1. **Data cleaning task:** The data was cleaned to remove certain columns where null count is greater than 40%. Empty values were replaced with NA.

2. **Creating dummy variables:** Dummy variables were created for categorical inputs.

3. **Train-Test split of the data**: The data was split into 70% training data and 30% test data.

4. **Building the model:** Features were selected using RFE. 15 features were selected

5. **Model Evaluation**: The model was evaluated using a classification report. The training data's sensitivity achieved was 69.54% and specificity of 88.26%.

6. **Predictions on train data**: Prediction was done on the test data frame and with a threshold of 0.4.

7. **Precision-recall curve**: Using a cutoff of 0.35, the precision value found, the model achieved a precision value of 75.46% and recall value of 75.85%

8. **Most important parameters**: The 10 variables that matter most:

TotalVisits - 5.727639
Total Time Spent on Website - 4.614182
Lead Origin_lead add form - 3.756959
What is your current occupation_working professional - 3.655520
Lead Source_welingak website - 2.582793
Last Notable Activity_unreachable - 1.806575
Lead Source_olark chat - 1.578001
Last Activity_sms sent - 1.261604
What is your current occupation_student - 1.221821
What is your current occupation_unemployed – 1.139414

9. **Conclusion** – The above parameters were determined to be the most important parameters according to the model. The model achieved a precision value of **75.46**% and recall value of **75.85%**