Lead Scoring Case Study Summary

The analysis was done for identifying the potential leads for a company X Education which offers online courses for working professionals. Checking the dataframe gave the information on how the potential customers visit the site, the time spent by them etc. and how many of them end up buying the courses.

The analysis went through the following steps:

1. **Data cleaning:** The columns with more than 3000 null values, more than 50% entries as ‘Select’, with almost all values as ‘No’ and some not so relevant columns like ‘City’, ‘Country’ were dropped. After applying all the criteria to remove null values and the unwanted columns and rows, 69% of the rows in the dataset were still retained.
2. **EDA:** A lot of entries in the categorical columns were irrelevant. The numerical variables showed proper relationships.
3. **Dummy variables:** Dummy variables were created and their respective original columns dropped.
4. **Train-test split:** Dataframe split into the train set with 70% and the test set with 30%. Feature scaling was done using MinMaxScaler.
5. **Model Building:** RFE was done to get the top 20 variables. Features were then eliminated based on the VIF (>5) and high p-values. After elimination of features one-by-one, the VIF of the features remained below 5 and the p-values below 0.05.
6. **Model Evaluation:** A confusion matrix was created. Area under the ROC curve was 87% and the optimal cutoff was 0.42. Using this cutoff, accuracy, sensitivity and specificity were calculated whose values came to be around 78 %.
7. **Precision-Recall:** On the test dataset, the optimal cutoff of 0.44 was obtained which was then used to calculate precision and recall whose values were calculated to be 80.7% and 74.4% respectively.

The overall model accuracy was found to be around 78% after using 14 features which is close to the 80% target expected by the company.

The numerical features ‘Total Time Spent on Website’, ‘Total Number of Visits’ and ‘Page Views Per Visit’ were the most important leads which had the potential to convert people towards buying the courses. Among the categorical features, ‘Lead Origin_Lead Add Form’, ‘Lead Source_Olark Chat’, ‘Lead Source_Welingak Website’ and ‘Do Not Email_Yes’ had the highest potential leads in the decreasing order of their importance.

Based on this analysis after implementing a logistic regression model, X Education can identify the above features to increase their conversion rate to get visitors to buy their online courses.