Summary

This analysis is performed for X education to find ways to get more industry professionals to join their courses. They provided us a lot of information about the potential leads who visit the sites and how much time spent over there, how they reached them, and much more.

The steps we have proceeded with our Lead Scoring assignment are below:

1. Data Cleaning:

a. The first step was to prepare the dataset by removing redundant variables/features
b. There were not any redundant variables. After that, we proceed with null values. Removed columns having more than 30% null values.
c. After removing those columns, we found that some columns are having a label as ‘Select’ which means the customer has not chosen any answer to this question. So, we opted to replace this label with a ‘null value’. These are as good as null.
d. For the remaining missing values, we removed those particular rows which have missing values.
e. With EDA, we found some numerical columns have outliers. So, we select the Capping approach and cap the outliers with 5 and 95 percentiles.
f. We found some categorical features have low-frequency rows and we clubbed them together as a single group

2. Data Transformation:

a. We changed the binary variables with Yes/No values into 1/0
b. We changed the multicategory labels into dummy variables.

3. Data Preparation:

a. After transforming the data, we have done Feature Scaling: to scale all numerical columns in a single frame.
b. We check the correlation matrix of all the variables and make a note of the variables which have high correlation and can affect our analysis, to avoid multicollinearity in our model.

c. We split the dataset into train and test datasets.

4. Model Building:

a. We created our model by using Recursive Feature Elimination (RFE) with counts 20 and 15 and chose our final model that starts with 15 variables and we finalize 10 variables that have more stability and accuracy than the other.
b. For the final model, we checked the optimal probability cutoff and checked Accuracy, Sensitivity, and Specificity.
c. We found a convergent point and we took that point as the cutoff point and predicted all our final outcomes.
d. We assigned a lead score to each lead between 0 and 100 so that it can help the company to target potential leads. A higher score -> **'Hot Leads'** and lower score -> **'Cold Leads'**.
e. We checked the Precision and Recall with Accuracy, Sensitivity, and Specificity. For our final model and tradeoffs.
f. After training, we made predictions on the test dataset and record all the predicted values.
g. We evaluated the model on the test dataset by checking the Accuracy, Sensitivity, and Specificity to find that model was trained good or bad.
h. We found that all those parameters came in an acceptable range for the test dataset.

5. Conclusion: Learning gathered

a. In business terms, our model is stable in terms of all parameters. It will surely give benefits to the company in future.
b. Top features with good conversion rates are:
    i. **Total Time Spent On Website**
    ii. **Lead Origin_Lead Add Form**
    iii. **Lead Source_Olark Chat**
    iv. **Last Activity_SMS Sent**
c. Lead Score calculated on Train dataset is 79.6% conversion rate and 80% conversion rate on Test Dataset. Hence, overall, this model seems good.