Summary of Lead Score Case Study

1. **Problem Statement:**

To Increase the conversion rate of an education company X Education from 30% to 80%.

As a data analyst we are supposed to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

To Perform the entire analysis and building predictive lead score model using the provided data below are the steps we took:

2. **Reading and Cleaning data:**
   First, I read and understood the data and their definition. then for cleaning I started with the null values handling then dropping the irrelevant fields and single value fields Also we changed the null values as ‘no info available’ so that we don’t lose available info, post that also categorized Country field in ‘India’, ‘outside India’ and ‘no info available’. Also dropped the unique value field which was not relevant. .

3. **EDA:**
   In EDA we extensively look for each field, visualized each of them summarized the insight and took appropriate decision. Compared categories against the converted.
   I found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.

4. **Data Preparation for Model Building**

a. **Dummy Variables:**
   Created The dummy variables were created, and later originals were removed.

b. **Train-Test split:**
   The split was done at 70% and 30% for train and test data, respectively.

c. **Feature Scaling:** done feature scaling using the Standard Scaler

5. **Model Building on train and test set:**
   Firstly, we did RFE which was done to attain the top 15 relevant variables. Later the rest of thevariables were removed manually depending on the VIF values and p-value (Thevariables with VIF < 2 and p-value < 0.05 were kept) also tried multiple cut off to optimize performance of the model.

6. **Model Evaluation:**
   A confusion matrix was made on cutoff 0.50. Later, the optimum cut off value 0.45 using ROC curve which was covering almost 92% area, was used to find the accuracy, sensitivity and specificity which came to be >82% each.

7. **Prediction:**
   Prediction was done on the test data initially frame and with an optimum cut off as 0.45 accuracy, sensitivity, and specificity of 80%.

8. **Precision – Recall:**
   This method was also used to recheck and a cut off 0.45 was found with Precision around 82% and recall around 83% on the test data frame.

Conclusion & Findings:

As per our model variable that are important in converting a lead are:

1. Where Lead Origin is 'lead add form'

2. Where Lead Source are:
   - 'olark chat'
   - 'welingak website'

3. Occupation is 'working professional'

4. Last Notable activity is 'Modified'

5. Last Activity is 'SMS Sent'

6. Lead Quality is 'Not Sure' and 'might be' (not sure are the cases where sales team did not fill anything)

7. Also some less important variables are that are behaviorally looking off:
   - Last Activity email bounced
   - Lead Quality worst
   - Last Notable Activity unreachable

We can further work on fine tuning for this but for its looking convincing.

THANK YOU 😊