Summary Report

The analysis is done for X Education to find ways to get more industry professionals to join their courses. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach/Steps to resolve the problem:

1.Reading and understanding the data:
The data provided gives a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

2.Cleaning data:
The data was clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

3.EDA:
A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

4.Dummy Variables:
-We created dummy variables for the categorical variables.
-Removed all the repeated and redundant variables.

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

6.Model Building:
-Firstly, RFE was done to attain the top 15 relevant variables.

- Rest of the variables were removed manually depending on the VIF values and p-value i.e. the variables with VIF < 5 and p-value < 0.05 were kept.
- we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good.
-For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
-We then plotted the ROC curve for the features and the curve is pretty decent with an area coverage of 86% which further solidified the model.
- checked if 80% cases are correctly predicted based on the converted column.
-checked precision and recall with accuracy, sensitivity and specificity for our final model on train set.
-Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.4.
-Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81.27%; Sensitivity= 84.22%; Specificity= 76.43%.

The variables that mattered the most in the potential buyers are in descending order :

1. **Lead Origin_Lead Add Form**
2. **Current Occupation_Unemployed**
3. **Current Occupation_Others**
4. **Lead Source_Welingak Website**
5. **Last Activity_Email Bounced**
6. **Last Notable Activity_SMS Sent**
7. **Last Notable Activity_Others**
8. **Lead Source_Olark Chat**
9. **Total Time Spent on Website**

Conclusion:

We got around 1% difference on train and test data's performance metrics.This implies that our final model didn't overfit training data and is performing well.

High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.

Depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

Recommendations:

The Team should focus on following variables:

1. **Total Time Spent on Website**¶
    - From The Boxplot we can see that number of constumers who have converted spent more time on the website.so try to make them engage in the website and increase their chances of getting converted.
    - Try to give them offers exclusive on websites so they gets attracted to it.
2. **Lead Source**
    - ***Olark Chat:***
        - We can see that olark chat has majority of lead conversions in terms of count, so focus on increasing the lead conversions to this categories.
    - ***Welingak Website:***
        - Welingak website has higher number of lead conversion rate but count is low.so you can improve and learn from here that might help you in improving other categories.
3. **Lead Origin**
    - ***Lead Add Form:***
        - Lead add form has higher number of lead conversion rate so you can improve and learn from here that might help you in improving other categories.
4. **Current Occupation**
    - ***Unemployed:***
        - Unemployed has highest number of conversions by count but conversion rate is low so focus on increasing the conversion rate on unemployed as it is obvious that unemployed will be most interested in getting a new job by learning something new.
    - ***Others:***
        - Others contains Student,Housewife and Businessmen so focus on them more, try to increase their conversion score to more then 35 which are potential customers.
5. **Last Activity**
    - ***Olark Chat Conversation:***
        - **Need to focus more on Olark chat as the conversion rate is poor.**
    - ***Email Bounced:***
        - **The conversion rate is not good seems to be very poor so focus here also.**

- *Email Opened:*
    - **Most leads are generated by email opened so this is an important category that you should focus on.Although having most conversion counts by number the conversion rate is not good.**

6. Last Notable Activity
    - *SMS Sent:*
        - **SMS sent has good conversion rate and so there is chance of improvance more.**
    - *Others:*
        - **Others has many categories such as 'Unreachable','Unsubscribed','Email Bounced','Had a Phone Conversation', 'Approached upfront','View in browser link Clicked','Email Received','Email Marked Spam','Form Submitted on Website', 'Resubscribed to emails'.**
        - **Since the lead count and converted count is less try on increasing the number as your leads score depends on this.**
7. Specialization
    - *Not specified:*
        - **As most of the customers has opted to not provide their specialization so avoid using this feature. Focus on leads having conversion score more then 35 which are potential leads also known as 'Hot leads'**