Lead Scoring Case Study Summary

This Analysis is done on X Education and to find ways to get more industry professionals to join their courses. The Basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and conversion rate.

The Following are the steps performed.

Step 1: Reading And Understanding the Data.

Read and inspected the data.

Step 2 : Data Cleaning:

- First step to clean the dataset we chose to drop the variables having unique values.
- Then there were few columns with the Value 'Select' which means the leads did not choose any given option. So changed those values to the Null Values.
- We Dropped the Columns having Null values greater than 30%.
- Next, We removed the imbalanced and Redundant Variables. This Step also included imputing the missing values as and where required. The Outliers were identified and removed.

Step 3: Data Transformation:

- Change the binary variables into 0 and 1.

Step 4: Dummy Variables Creation:

- Created the dummy variables for categorical Variables.
- Removed all the Repeated And Redundant Variables.

Step 5: Test Train Split:

- Data was split into Train and Test Dataset with a proportion of 70 – 30 % respectively.

Step 6 : Feature Scaling:

- Then We Used the Standard Scaler to scale the original numerical Variables.

Step 7 : Model Building & Model Evaluation:

- Using the Recursive Feature Elimination, we went ahead and select the top 15 important features.
- Using the Statistics generated, we recursively tried looking at P- Values in order to select the most significant values that should be present and dropped the insignificant values.
- For our final model we checked the optimal probability cut-off by finding points and checking the accuracy, sensitivity, specificity.
- We then plot the ROC curve for the features and the curve came out to be pretty decent with an area coverage of 88 % which further solidified of the model.
- Then checked if 80% cases are correctly predicted based on the converted column.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next based on the Precision-Recall Trade-off, we got the cut off value of approximately 0.34.
- Then We implemented the learnings to the test model and calculated the conversion probability based on Sensitivity and Specificity metrics and found out the accuracy value to be 80.05%, Sensitivity = 81.44% and Specificity = 79.16%.

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 81% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute more towards the probability of lead getting converted are:
    1. **Lead Origin_Lead Add Form**
    2. **Last Notable Activity_Other_Notable_activity**
    3. **Last Notable Activity_SMS Sent**