Lead Scoring Case Study Summary

As a part of Lead Scoring case study, we have been given with a data from X education company that sells online courses. From the data given we need to analyze and identify the Hot Leads, i.e, the leads that gets mostly converted.
As mentioned the conversion being 30%, we need to analyze the data and come up with a model which can make predictions as such the conversion should be 80%.

Following describes the steps that has been followed to achieve the result.

1. Read and inspected the data.

2. **Data Cleaning:**
   - Dropped the columns with unique values
   - The columns with 'Select' value have been nullified since it indicates that people didn't choose any option.
   - Dropped the columns having null values greater than 40%.
   - Outliers are examined and dealt with missing values.

3. Identifying the potential data columns which can factor in for accurate prediction
4. Identifying the relationship and distribution of column data using graphs
5. Removing the outliers in numerical variables
6. Plotting heat map to see the correlations
7. Created dummy variables for all the categorical variables
8. Split the data into train and test sets of 70% and 30% propotions respectively

9. Model Construction

- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
- Using the statistics generated, we recursively tried looking at the 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 and specificity.
- We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 97% which further solidified the 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 and Recall trade-off, we got a cut off value of approximately 0.3.
- 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 92.66%%; Sensitivity= 91.58%; Specificity= 93.17%

10. Conclusion:

Lead scores >0.35 have a higher conversion rate and a 91% model accuracy score.