REPORT

MODEL HAS BEEN BUILT FOR X EDUCATION COMPANY USING THE FOLLOWING PROCESS

1. **Exploratory Data Analysis:**
   - First, we removed unnecessary columns which dealt with ID and Lead number, which were all unique and wouldn’t have helped us with the model
   - Secondly, we replaced ‘Select’ with null values, as it meant that user didn’t select any options.
   - Then we removed columns which had more than 40% missing values.
   - After that we dealt with columns having missing value more than 10% by mean(), median() or mode() method.
   - Then we saw the contents of columns, of which some turned out to have excessive asymmetricity, we removed such columns too.
   - We also merged certain contents into certain categories for easy analysis.
   - Moreover, we worked on numerical variable, outliers and dummy variables.
   - Finally, we plotted various graph to understand the distribution

2. **Data Preparation :**
   - We transformed binary columns such as ‘Do Not Email' & 'A free copy of Mastering The Interview' to ‘0’ & ‘1’ from ‘No’ & ‘Yes’
   - Then we created dummy variables

3. **Test-train Split :**
   - We splitted the data into Test-train in the proportion of 35% and 65%.

4. **Model Building & Evaluation**
   - We selected 10 variables using RFE
   - Then we removed variables depending on the VIF values.
   - At first, confusion matrix was created with 0.5 cutoff and accuracy came at about 78%

- After optimizing cutoff at 0.3, we got accuracy of 78%, sensitivity of 77% and specificity of 78% on training set and 77%, 75% and 78% on test set respectively.
- We also checked precision- recall rate at 0.3 cutoff, and we got 76% and 62%
- Finally we did precision-recall tradeoff at 0.304, which gaves us accuracy of 78%, precison and recall of 69% and 76% on training set. And on test set, we got accuracy of 77%, precision of 70% and recall of 75%

So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be **0.30**
&
If we go with Precision – Recall Evaluation the optimal cut off value would be **0.304**

CONCLUSION

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.

The variables that mattered the most in the potential buyers are (in descending order):
Total Time Spent on Website | Coefficient: 3.35
Lead Origin_Lead Add Form | Coefficient: 2.78
Lead Origin_Lead Import | Coefficient: -2.21
Page Views Per Visit | Coefficient: -1.96
Do Not Email | Coefficient: -1.64
Lead Source_Olark Chat | Coefficient: -1.25
What is your current occupation_Working Professional | Coefficient: 1.17
Lead Origin_Landing Page Submission | Coefficient: -0.75
Lead Source_Reference | Coefficient: -0.69
A free copy of Mastering The Interview | Coefficient: -0.18
Lead Source_Organic Search | Coefficient: -0.16
TotalVisits | Coefficient: -0.15

-