Summary – Lead Scoring Case Study

Problem Statement – The lead conversion rate of X Education Private Limited, an education company was very poor @30%.

Target – Maximize lead conversion to 80% conversion

Process – Data Modelling by assigning lead score

Steps :

1. **Inspecting the Data** – Necessary libraries were imported and the dataset 'Leads.csv' was inspected using head, info and describe functions.

2. **Data Cleaning and EDA** – Prospect ID was dropped as the unique identifier with Lead number sufficing. All columns with > 40% null values were dropped in the next step, however still some columns remained with substantial % of null values. The null values were imputed with relevant ones for columns such as 'Tags', 'Lead Profile' etc. Other columns like 'Country', 'City' were dropped due to presence of many null values.

3. **Univariate Analysis** – Categorical variables such as Lead Source, Lead Origin etc. were plotted against the target variable 'Converted' and important insights were drawn specific to each of these variables. Numerical variables such as Lead Number, Total Visits etc. were plotted against each other and target variable 'Converted', however not yielding significant correlation in the heatmap.

4. **Data Preparation** – The data was further prepared for modelling in this step.

   a) **Converting binary variables (Yes/No) to 1,0** – Binary variables in columns 'Do not email' and 'a free copy of mastering interview column' were converted to 1 & 0.
   b) **Dummy Variable** – Dummy Variables were created for all categorical columns
   c) **Outlier Treatment** – Variables 'Total Visits' and 'Pages view per visit' were found to have outliers, which were treated by capping values at 99%.

5. **Train – Test Split** – The data is then split into test and train sets in the 70:30 ratio.
6. **Feature Scaling** – The features were scaled using StandardScaler() module to make their values comparable.
7. **Correlation** – The correlation achieved through heatmap did not yield desirable results due to large number of variables
8. **Model Building** – The first model was built using logistic regression module, however improvement was needed.
9. **Feature Selection using RFE** – Using RFE module from sklearn library, 20 features were selected at random and a second model was built using statsmodel library. The variables with p value > 0.05 in the model summary were dropped. A third model with remaining features, yielded p values <0.05. The model accuracy was 90.32%.

10. **Checking for VIFs** – Few variables such as 'Last notable Activity', 'Lead Number' had VIFs > 5 were dropped, one at a time, starting with 'Last Notable Activity'.
11. **Fourth Model** – The 4th model built achieved an accuracy of 90.16% (not dropped much) with VIFs under 5 for all variables. The metrics at this moment were:

Accuracy – 90.2%	**Sensitivity** – 83.2%	**Specificity** – 94.5%
False Positive Rate – 5.54%	**Positive Predicted Value** – 90.2%	**Negative Predicted Value** – 90.13%

12. **ROC Curve and Optimal cutoff point** – ROC curve was plotted, and optimal cutoff point was calculated as 0.3. The accuracy and other metrics were recalculated**.**
13. **Precision and Recall** – The precision and recall came out to be 84.96% and 88.48% respectively.
14. **Model Evaluation** - The model made predictions on the test set with great accuracy of 90.4%
15. **Lead Score** - The lead score was assigned to all values.
16. **Final Observations**

Comparison: Train vs Test:

	Accuracy	Sensitivity	Specificity
Train Set	89.6%	87.5%	90.3%
Test Set	90.4%	88.5%	92.7%

2) The model seems to predict the conversion rate quite well.

3) The top variables to focus on (as per the coeffients of model summary and heatmap)

- Total Time spent on the website
- Lead Origin – Lead add form
- Last Activity – When last activity is SMS sent

17. **Final Recommendation**: X Education can maximize conversion rate by targeting leads with the below specifics

Top 3 variables

Total Time spent on the website

Lead Origin – Lead add form

Last Activity – When last activity is SMS sent

Other variables

Lead Source – References, Wellingak Website, Google and Direct Traffic
Lead Occupation – Working professionals as the current occupation
Specialization – Management, Banking Investment & Insurance and Business Administration have almost 40-50% lead conversion rate. Leads from these areas can be targeted