Summary Report

Problem Statement:

- Company X Education receives leads where potential customers fill forms via multiple channels requesting information about certain courses being provided by the company.
- Only about 30% of the leads get converted to actual customers.

Solution Methodology:

- Build a logistic regression model which will help determine the most promising leads assigning a lead score from 0 to 100 to every lead.
- Expected conversion rate using model to be around 80%.

Steps performed for analysis:

- **Reading and understanding data**
    i. Variables with null values were determined.

- **Data Preparation**
    1. **Variable treatment:**
        1. Converted values marked 'Select' to NaN.
        2. Checked for contribution of each category value towards conversion: Following variable values appear to have a higher conversion rate –

Variable	High conversion values	Recommendations
Lead Origin	Lead Add Form	Spend more on advertising using Lead Form
Lead Source	'Welingak Website' and 'Reference'	Consider advertising more on 'Welingak Website'. Ask existing customers to provide references.
Last Activity	SMS Sent	Follow up using SMS
Last Notable Activity	SMS Sent	Follow up using SMS
What is your current Occupation	Working Professional	Target more working professionals
Tags	Will revert after reading the email, Closed by Horizzon	Follow up post sending email

        3. Checked for variables with highly skewed data (More than 85% one value): This indicates very less variance in data and is not useful for model building.
        4. Checked for variables with values distributed such that some values had higher values but some values were less than 10% of distribution: Combined the minimal values as new category - Others.
    2. **Missing Value imputation:**
        i. Imputed missing values for numeric and categorical columns with <2% missing data by median and mode respectively.

ii. Variables with missing values around 35% were not imputed and left as NaN since imputing median values could skew data here.

3. **Binary variables and dummy variables were created, and repeated variables dropped.**
4. **Outlier treatment:**
   - Outliers were hard capped to 95% on the higher side only.
5. **Correlation:**
   - Created heat map and pair plots to look for correlation.

- **Model Building:**
  1. **Test/Train Split:**
     - Split data 70/30 to train and test.
  2. **Feature Scaling:**
     - Used Standard Scaler to scale numeric columns.
  3. **Feature Selection:**
     - Used RFE for feature selection.
     - Checking p-value and correlation (VIF). Dropping highly correlated features
  4. **Model Metrics:**
     - Determined model Accuracy, other metrics using confusion matrix and AUC using ROC curve.
  5. **Optimal Threshold:**
     - Rebuilt model as per optimal threshold: 0.3.

- **Model Evaluation:**
  - Ran model on test data and calculated metrics. Metrics in train and test data were observed to be similar.

Train dataset

	Metric	Value
0	accuracy	0.878324
1	sensi/TPR/Recall/HitRate	0.875101
2	speci/TNR	0.880310
3	FPR	0.119690
4	FNR	0.124899
5	PositivePredictiveValue/Precision	0.818354
6	NegativePredictiveValue	0.919603

Test dataset

	Metric	Value
0	accuracy	0.874459
1	sensi/TPR/Recall/HitRate	0.874886
2	speci/TNR	0.874180
3	FPR	0.125820
4	FNR	0.125114
5	PositivePredictiveValue/Precision	0.819504
6	NegativePredictiveValue	0.914535

- As per model built, the probabilities were converted to Lead Scores by multiplying by 100.
- Scores from 0 to 100 indicate whether that higher the score the hotter the lead.
- Important features and values observed were:

- Tags_Will revert after reading the email
- Lead Origin_Others
- Last Notable Activity_SMS Sent