Summary

Problem Statement:

X Education is an online education company targeting industry professionals through various marketing channels and collect leads. Their lead conversion rate is low (~ 30%). The company wants to identify "Hot Leads" (highest conversion potential) to increase conversion rates. A lead scoring model is required that assigns scores to leads, for sales team to prioritize high-scoring leads. The CEO aims for target lead conversion rate of approximately 80%.

Data Cleaning:

- Columns with >40% nulls were dropped.
- 'Select' values in columns were replaced with NaN.
- Identifier columns (e.g., Prospect ID) were dropped.
- Highly skewed columns with one category were dropped.
- Combined categories with very low row percentages as it does not make sense to create dummies for such values.
- Imputing column with least missing values percentage.
- Dropped rows with null values.
- Renamed long column names to short and appropriate ones.
- Checked rows retained (98.2%) after completing data cleaning steps.

EDA:

- Checked and only ~ 38% leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables.
- 'Lead Origin', 'Lead Source', 'Current_occupation' etc. impacts target variable.
- Numerical variable 'Time spend on website' shows positive impact on lead conversion.
- No insights could be related to lead conversion from 'Total Visits', 'Free_Copy' columns.

Data Preparation:

- Created dummy variables for categorical variables e.g., Lead Origin.
- Splitting Train & Test Sets in 70:30 ratio.
- Scaled numeric variables using Standardization.
- Checked correlated variables using heatmap.

Model Building:

- Used RFE to reduce variables from 45 to 15(coarse tuning)
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05 and VIF > 5.
- Total 3 models were built to come up at the final model (Logm3 with 13 variables) which was stable with (p-values < 0.05 and VIF < 5).

Model Evaluation:

- Created confusion matrix to calculate specificity (~ 88.70%), sensitivity (~74.08%) . Accuracy was 81.62%.
- Drew ROC curve (using accuracy, specificity, and sensitivity) plot to come up at cutoff point of 0.38. Using the new cutoff, performance metrics were recalculated.
- Precision-recall curve was drawn to get cutoff of 0.44. Accuracy on train set was 81.62%. Specificity and sensitivity dropped by few decimal points.
- Lead score was assigned to train data using 0.44 as cut off.

Making Predictions on Test Data:

- Numeric features were scaled, and lead score was predicted using final model.
- Accuracy on test set was 82.07% with sensitivity and specificity close to 74%
- Lead score was assigned.
- The conversion rate with the model is 95% for lead_score>=95, and 87.02% for lead_scores >= 80
- Top 3 features are:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current_occupation_Working Professional

Recommendations:

- The company should target the leads originated from Landing Page as they have higher chances of conversion.
- Company should focus on people who are:
    - Unemployed, Working Professional and Students.
    - Active on Email Opened.
- Total Time Spent on Website is an important factor impacting lead scores.
- 447 leads fetched using model have a high chance of getting converted if we target lead_score >= 80.