Summary

Our objective was to build a model to identify potential leads so as to increase the X Education company conversion rate.

Step 1: Data Understanding and Cleaning

- Inspect the data, look at data structure, and go through the variables.
- Columns with high missing values, (>40%), are dropped to maintain data integrity.
- We have a label ‘Select’, consider it as a null value.
- Dropped columns that did not seem much of relevance for e.g. location variables, unique identifiers; also dropped rows with missing values less than 5%.
- Merged and grouped few labels and also dropped the columns that have only single label since they are highly imbalanced.
- Check for outliers and remove the extreme datapoints that could skew the model.

Step 2: Exploratory Data Analysis (EDA)

- 37.9% of leads are converted.
- Distribution of TotalVisits variable showed a right-skewed distribution and for Lead Origin variable, 'Landing Page Submission' turned out to be the most common origin of leads.
- For columns with 2 unique values the ones with one label completely dominating over other were dropped. Also visualized that maximum leads being converted were from the ‘Management’ Specialization.
- Box plotted sources against the total visits, ‘Organic Search’ had several visits with a good conversion rate. For Total Time Spent on Website by Occupation analysis, category ‘Others’ spend a lot of time on website, but still have a low conversion rate.
- From the heatmap only Total Time Spent on Website showed moderate positive correlation with target variable.

Step 3: Data Preparation

- Create dummy variables for categorical columns, drop one category to avoid multicollinearity and that is reference category.

Step 4: Model Building

- Split the data, then scale using StandardScalar() and fit and transform the training set.
- Study the correlation, using .corr(), since too many features let RFE select 15 most relevant features.

- Built a logistic regression model using GLM to predict lead conversion, remove the ones with high p-value i.e. above 0.05 and check if VIF value is above 5, drop those variables to avoid multicollinearity.
- Predict() is used on fitted model to predict probabilities
- Randomly choose threshold of 0.5 to classify.
- Created confusion matrix, got accuracy of 81.76%, but the ones predicted wrong are left out.

Step 5: Model Evaluation

- Got 70.24% sensitivity i.e. model correctly identifies 70.24% of actual positives and specificity as 88.88% means model correctly identifies 88.88% of actual negative cases.
- Plotted the ROC curve and it depicted high AUC.
- Find optimal threshold, at different thresholds the plotted lines intersected at 0.37.
- Calculated the metrics again at 0.37: Accuracy was 81.11%, Sensitivity: 78.97% and Specificity: 82.53%.
- Plot Precision-Recall curve, and at 0.42 threshold value there is balance between both.

Step 6: Predict on Test set

- Scale the same parameters and transform to the test set.
- We make prediction on the subset of features [col] that were used during model training, set our threshold of 0.42 to classify.
- Calculated the metrics: Accuracy: 80.28%, Sensitivity: 72.46%, Specificity: 84.95%, Precision: 74.21%
- Added a new column to show the scores of leads.