Summary

The analysis is done for X education company that sells online courses to the industry professionals. The main motive of the company is to identify the most potential leads that are likely to join their courses. The leads.csv data has important information regarding how much time they spend on the website, how many times they visited the website, how they reached the site, etc.

Approach:

Following series of steps were followed:

1. EDA

This involved getting the basic understanding of the data.

- Invalid values: The invalid values like 'Select' were replaced with NaN
- Missing value: Missing values were handled. The columns with more than 40% missing values were dropped.
- Merge small frequency levels: In categorical columns, small frequency levels, have been merged to form new level and the missing values were replaced with mode or 'Not Specified'.
- Outlier handling: The outliers in numeric columns were handled.
- Check imbalance: Target variable, converted was checked for data imbalance
- Visualizations: Count plots, boxplots, pair plots and correlation heatmap were visualized for converted and non-converted leads.

2. Data Preparation
    - Binary Encoding: The binary variables are encoded from Yes and No to 1 and 0 respectively.
    - Create Dummy variables: The categorical variables are dummified.
    - Train-Test split: The data is split into train and test datasets in 70-30 ratio.
    - Standardisation: The numerical columns are scaled using standard scaler.
    - Determining X and Y: Dataset is split into X and Y

3. Building the model (using Train set)

- Used RFE to find top 16 columns impacting the lead conversion
- Used statsmodels to build a logistic regression model and used p-values and VIF to eliminate the independent features. (The variables with VIF > 5 or p-value > 0.05 were removed)
- Predicted the probability of lead being converted.
- Used 0.5 as default threshold and calculated metrics like Accuracy, Sensitivity, Precision, etc.
- Plotted ROC curve
- Determine Optimal threshold by plotting Sensitivity, Specificity and Accuracy for thresholds in the range 0 to 1
- Found optimal threshold and calculated metrics like Accuracy, Sensitivity, Precision, etc.

4. Making Predictions (on test set)
    - Standardisation: The numerical columns are scaled using standard scaler.
    - Determining X and Y: Dataset is split into X and Y
    - Predict probability of lead being converted and assigned lead score to each customer
    - Using optimal threshold, predict converted and calculated metrics like Accuracy, Sensitivity, Precision, etc.

Learnings:

Company should focus on leads with

- Current status of lead (tag): Closed by Horizzon, Lost to EINS, Will revert after reading the mail
- When the lead source was Welingak website or Olark chat
- When the lead origin is Lead add form.
- When the last activity was SMS Sent
- Total time spent on the website.

There is no need to focus on leads with

- Current status of lead (tag): Switched off, Ringing, already a student, Interested in other courses
- Last notable activity: Olark chat conversation, Activity Modified

If company wants Aggressive lead conversion, they can target all leads with Lead Score greater than 30. If the company wants to minimize the calls, they can just focus on the leads that have the Lead Score greater than 70.