Leads Scoring Case Study Summary

Below are the steps on how we have proceeded with this case study:

1. Data Cleaning:

a. The first step was to clean the dataset where we chose to remove the redundant variables/features.
b. After removing the redundant columns, we found that some columns are having label as 'Select' which means the customer has chosen not to answer this question. The ideal value to replace this label would be null value as the customer has not opted for any option. Hence, we changed those labels from 'Select' to null values.
c. Removed columns having more than 40% null values.
d. For remaining missing values, we have imputed values with median values for continuous columns and with mode for categorical columns.
e. We found for one column had two identical label names in different formats (capital letter and small letter). We fixed this issue by changing the labels names into one format.

2. Data Transformation:

a. Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.
b. Checked the outliers and created bins for them.
c. Removed all the redundant and repeated columns.
d. After this, we plot a heatmap to check the correlations among the variables and dropped variables with multicollinearity.

3. Model Building:

a. We split the Train-Test data in 30:70 ratio and created our model with RFE count 25 and compared the model evaluation score like AUC and chose our final model.

b. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity. Optimal probability cutoff came out to be 0.3.

c. We checked the precision and recall with accuracy, sensitivity and specificity for our final model.

d. We made prediction in test set and predicted value was recoded.

e. We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is:

- We found the score of accuracy and sensitivity from our final test model is in acceptable range.
- We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

4. Conclusion:

a. Our Logistic Regression Model is decent and accurate enough, with near 77% Accuracy and 83% Sensitivity.

b. X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on Welingak website.
- Approach to Working Professionals who want to gain extra knowledge, since this helps in higher conversion.
- Focus on lead form on the website.
- Add quality contents on website so that user will spend more time exploring information about courses.