Lead Scoring Case Study Summary

Problem Statement

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary

For the given problem we used Logistic regression on given dataset and followed below steps:

1. **Data loading and understanding of data:**
    - We loaded data and an initial inspection was done to understand the basic characteristics of the data such as data type, size, shape, distribution, missing values, and so on.
2. **Data cleaning:**
    - We dropped variables which had single value (single value features), also dropped columns which had null percent greater than 50%.
    - Imputing of values were done for categorical columns with mode and numerical data types with median.
    - Removed couple of columns that had unique values which was not necessary for the analysis.
    - Outliers were treated by using capping method base don the distribution of data in its respective columns.
3. **Exploratory Data Analysis (EDA):**
    - After data cleaning and outlier treatment was done we performed EDA as this is the process of visualizing and analyzing the data further to gain insights and understand the underlying patterns and relationships between variables.
    - This step helped us to cross verify if there were any outliers still present, correlation between variables, and other patterns that was useful in building the model.
4. **Data pre-processing and feature scaling:**

- In this step we transformed data into a format that is suitable for model building. scaling of numerical variables was done using standardization method, converted Nominal categorical variables to Numerical variables using one hot encoding and ordinal variables with label encoding.
- This step was done as it is essential for improving the performance of the model.

5. **Model Building:**
    - Used RFE for feature selection to get top 30 significant variables. We split data into Train and Test sets with 70:30 ratio respectively.
    - Once the data preparation as mentioned in above step is done, we did model building by manually removing variables depending on p-value and VIF i.e., if p-value > 0.05 and VIF > 5 we dropped those columns.
    - Model building is done on training data and this model will be later used in the next step to validate using test data to ensure it is performing well.
6. **Model Evaluation:**
    - In this step, the performance of the model was evaluated on a test dataset.
    - Confusion matrix was created using which accuracy, sensitivity, specificity, false positive rate, positive prediction rate and negative prediction rate were calculated.
    - We got accuracy of test dataset as 82.97% and with train dataset it was 87.96% where optimal cut-off point was 0.36