Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals and it requires information to find the most promising leads in order to increase the lead conversation rate, The Company requires a model to be built in which lead score has to be assigned to each of the leads so that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance

Summary

1. Loading the Dataset

2. Data Understanding:
   We will understand each variable using Data Dictionary
   We will Check the structure of dataset
   We will check for other attributes like info (), describe (), shape etc.

3. Data cleaning:
   - Initial step was to find the duplicate values we found none
   - Next we checked for missing values in rows and kept a threshold of 35% since many columns were having more than 40% of missing data
   - Columns having missing values more than 35% were dropped since imputing these columns would certainly deprecate the data quality
   - For columns having less than 2% missing data we have dropped the rows, after which 98% of total data was retained.
   - Since all the numeric columns had less than 2% data so we have handled them during the above process.
   - The remaining categorical columns, which had null values between 15% - 30% of missing values but also were highly skewed data was also dropped.
   - Those Columns not having highly skewed data were imputed using the modal value of that column.
   - Columns having incorrect data type were also rectified.
   - Few categorical columns had a lot of categories, so those categories having very less value counts as compared to other categories, were clubbed together into a single category named ‘Other’.

4. Outlier treatment: There were 2 out of 4 numerical variables which had outliers, which were in huge amount, these were capped to 99.3rd percentile value.

5. Data Analysis using EDA:
   - EDA was done on all the remaining variables in the dataset in which we have checked for the imbalance in the 'Converted' variable
   - Uni-variate Analysis with respect to 'Converted' variable was performed on each variable for further understanding.

6. Data Preparation for modeling:
   - All the sales team generated columns were dropped as sales team generated columns come after lead generation
   - All variables having only 2 categories were converted to binary numeric variables with 1's and 0's
   - Dummy Variable were created for categorical columns which have more than 2 categories, of such variables the original variables were dropped
   - We then split the dateset using Train-Test Split and numeric features were scaled using Standard Scalar method.
   - Then we have used heatmap to check the correlation between variables

7. Data Modeling:

- Firstly, a basic model was created using all variables then using RFE, 18 Features were selected.
- Variables having high p-value and high VIF were dropped while creating a new model every time a feature was dropped, Our final model had 10 variables.

8. Model Evaluation:

- A data frame was created with Converted variable, probabilities of Converted Variable and a 'predicted' column containing 1's if probability was above 0.5 and 0's otherwise, here 0.5 was the assumed cutoff
- Then we plotted the ROC curve with respect to our final model and the area under curve came out to be 83% which further validated the model
- Then we plotted Sensitivity-Specificity and Accuracy plot from which the optimal cutoff came out to be 0.3 with Accuracy = 78.8%, Sensitivity =74.8%, Specificity = 81.2%
- Then using Precision-Recall trade off plot the optimal cutoff came out to be 0.35 with Accuracy = 79.5%, Precision = 72.8%, Recall = 72%
- A data frame was created for both cutoff's, 0.3 and 0.35 with Converted Variable, probabilities of Converted variable and a 'predicted' column containing 1's and 0's depending on the selected cutoff
- Finally, cutoffs of 0.3 was selected as the optimal cutoff because of sensitivity being high
- Then, predictions were made on the test dataset with cutoff being 0.3
- We achieved a final accuracy of 80.2%, sensitivity = 75.8% and specificity = 82.9%