SUMMARY REPORT

Problem Statement: An education company named X Education sells online courses to industry professionals. X Education company requires to select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

To solve the above problem, we have built logistic regression model on the given dataset. The steps that we followed are as below:

1. Imported required Python libraries like numpy, pandas, matplotlib.pyplot, seaborn, sklearn etc.
2. Loaded the dataset into jupyter notebook.
3. Checked the dataset and calculated column wise missing value percentages.
4. Removed the columns having more than 30% missing values.
5. Deleted severely imbalanced columns after imputing null values with mode.
6. Checking outliers for all the numerical variables by plotting boxplot.
7. Performed EDA to visualise the numerical variables by pairplots and heatmap and to visualize categorical variable by countplots.
8. Prepared the dataset for model building by creating dummy variables for all the categorical variables and converting binary categorical variable into 0 & 1 categories.
9. Splitted the dataset into train and test dataset.
10. Used standardscaler to scale the train data for a better model.
11. Built logistic regression model by using statsmodel GLM.
12. Selected 15 features by using RFE.
13. Created a dataframe with the actual converted flag and predicted probabilities.

14. Selected a cutoff of 0.5 to predict the probabilities of a lead to be converted i.e. if the prediction probability is more than 0.5 then the lead will be converted else not.
15. Checked VIFs of all features to tackle multicollinearity. Dropped features with high VIFs.
16. Reached to a final model where P values and VIF of all the variables are below 0.5 & 5 respectively.
17. Created confusion metrix where accuracy was 79%, sensitivity 68%, specificity 86%.
18. Plotted ROC curve to show the tradeoff between sensitivity and specificity.
19. Found optimal cutoff of 0.35 by calculating accuracy, sensitivity and specificity for various probability cutoffs.
20. Made predictions on test dataset and found accuracy 79.14%, sensitivity 79.77%, specificity 78.77%, precision 68.19%, recall 79.77% for test dataset.
21. Assigned lead score from 0 to 100 to identify 'hot leads' i.e. high score leads which means high probability of lead conversion.

Learnings from the case study:

1. Learnt how to deal with missing values and highly imbalanced columns.
2. How to find the optimal cutoff by calculating accuracy, sensitivity and specificity for various probability cutoffs.
3. How to handle multicollinearity.
4. How to interpret confusion matrix.
5. How to help company taking right strategies to maximise the lead conversion rate.