Summary

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

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%.

Summary:

Step1: Reading and Understanding the Data
Read the data from the CSV file provided and analyze the data

Step2: Data Cleansing

a. The variables that had high percentage of null values were dropped.
b. There were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to NaN.
c. Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.
d. All sales team generated variables were removed to avoid any ambiguity in final solution.

Step3: Data Transformation
Changed the binary variables into '0' and '1'

Step4: Analyze the Data
As part of the Exploratory Data Analysis, we tried to explore more on the provided data. We found some of the elements in categorical variables which were irrelevant along with further analysis of numeric values.

Step5: Dummy Variables
Dummy variables were created for the categorical variables. Removed all the repeated and redundant variables.

Step6: Train-Test Split
The next step was to divide the data into test and train sections with a proportion of 70-30% values.

Step7: Feature Rescaling
We used the Scaling feature to scale the original numerical variables. Then, using the stats model we created our initial model which would give us a complete statistical view of all the parameters of our model.

Step8: Model Building

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
b. Using the statistics generated, we recursively tried looking at the p-values in order to select the most significant values that should be present and dropped the insignificant values.
c. The VIF's for these variables were also found to be good.
d. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.
e. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
f. We then plot the ROC curve for the features and the curve came out be as expected.
g. We have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
h. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics.

Step9: Model Evaluation

A confusion matrix was derived. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80%.

Step10: Prediction

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of around 80%. Then, we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics.

Step11: Conclusion

Good value of sensitivity of our model will help to select the most promising leads. Features which contribute more towards the probability of a lead getting converted are:

a. Lead Origin_Lead Add Form
b. Lead Source_Welingak Website
c. What is your current occupation_Working Professional