SUMMARY REPORT

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 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 has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary

Step 1: Reading and Understanding Data

Read and inspected the data.

Step 2: Data Cleaning

- The first step to clean the dataset was to choose to drop the variables having unique values. Since the values in 'Prospect ID' and 'Lead Number' are unique ID of the customers, they were dropped.
- Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
- Next was to handle columns with a high percentage of missing values. So, we dropped columns with 45% or more missing values.
- Then we did an analysis on the categorical columns. Based on the column we did the following:

→ By checking the value counts on the column including null values, we replaced the null values with the mode of column 'City', 'What is your current occupation', 'What matters most to you in choosing a course'

→ In columns like 'Specialization', 'What matters most to you in choosing a course', 'Tags' where the null values were the highest, we replaced the null values with 'Not Specified'

→ In columns like ‘Lead Source’ and ’Last Activity’, we replaced null and low frequency values with ‘Others’.
→ After checking the percentage of null values in all columns we found a very insignificant percent. So, we dropped all rows with null values
→ We plotted a count plot to understand the spread of each categorical column with hue as ‘Converted’ column. We got a lot of insights from that.
→ We observed a lot of columns that did not have much variance in the data (majority of the data were in one category) So we decided to drop them.

- Then we performed analysis on numerical variables. By plotting heatmap, we checked correlations of numeric variables. For each column, we plotted a box plot to check for outliers. And based on the presence of outliers appropriate outlier treatment was done.

Step 3: Data Preparation

- We created dummy variables for the categorical variables. And dropped the original columns after dummy variable creation.
- The Train-Test split was done at 70% and 30% for train and test data respectively.
- We used the Standard Scaler to perform scaling on the original numerical variables

Step 4: Modelling

- We built a Logistic Regression model using Statsmodel and RFE technique to perform variable selection.
- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features
- Using the statistics generated, we recursively tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.
- Finally, we arrived at the 10 most significant variables. The VIF’s for these variables were also found to be good.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plotted the ROC curve for the features and the curve came out to be good with an area coverage of 0.97 which further solidified the model.

- We checked the precision and recall with accuracy of 90.81%, sensitivity of 92.05% and specificity 90.10% for our final model on train set.
- Next, based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.
- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 90.92%; Sensitivity= 91.41%; Specificity= 90.62%.

Conclusion

- The lead score calculated in the test set of data shows the conversion rate of 91% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- The 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

  I. Total Time Spent on Website

  II. What is your current occupation_Working Professional

  III. Last Activity_Converted to Lead