SUMMARY - Lead_Score_CaseStudy

Problem Statement:

X Education sells online courses to industry professionals and Students. 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%

The following are the steps used:

I. **Reading and Understanding Data**:
Read and analyze the data using describe and info

II. **Data Cleaning**:
We dropped the variables that had high percentage of NULL values in them (removed columns with more than 45% null values). 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.

III. **Exploratory data analysis:**
Then we started with the EDA of the data set to get a feel of how the data is oriented. In this step, there were columns that were identified to have only one value in all rows. These variables were dropped, as they would affect the outcome of logistic regression model as data in these columns was highly imbalanced.

IV. **Creating Dummy Variables**
we went on with creating dummy data for the categorical variables, then we divided the data set into test and train sections with a proportion of 70-30% values.

V. **Model Building:**
Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

VI. **Model Evaluation:**
A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.

VII. **Prediction:**
Prediction was done on the test data frame and with an optimum cut off as 0.30 with accuracy, sensitivity and specificity of above 90%.
Cutoff of 0.3 was decided based on plot of accuracy, sensitivity and specificity for various probabilities.

VIII. **Computing the Precision and Recall metrics:**
we also found out the Precision and Recall metrics values came out to be 87.8% and 91.3% respectively on the train data set.

IX. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 90% each.

X. Making Predictions on Test Set:

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 92.78%; Sensitivity=91.98%; Specificity= 93.26%.

It was found that the variables that mattered the most in the potential buyers are:

1. Tags:
   a. Closed by Horizzon
   b. Lost to EINS
   c. Will revert after reading the e-mail
2. The total time spend on the Website.
3. Total number of visits.
4. When the lead source was:
   a. Welingak website
   b. Direct traffic
   c. Organic search
5. When the last activity was:
   a. SMS
   b. Olark chat conversation

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.