Summary Report of Lead Score case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. Company gets leads through different sources. Currently only 30% of leads get converted while others do not get converted.

Analysis done on the leads to increase the conversion rate to 80%. Built a logistic regression machine learning model to determine the leads which had more probability of conversion. Also identified the features which can increase the conversion rate more than 80%.

Source Files provided:

1. Leads.csv
2. Leads Data Dictionary.xlsx

Machine Learning steps Performed:

1. Clean Data preparation:
   Given input data contains 9240 rows and 37 columns. After analyzing the data, we found many missing values and select in many category features. Converted select as null values and removed all rows with null values. Removed few columns which had high skewed data. Imputed nan values with most repeated values. After data cleaning, we retained 98% of data.

2. Exploratory Data Analysis EDA:
   Checked for any duplicate values. Done univariate analysis and bivariate analysis, made the inference for all the features. After univariate and bivariate analysis, we came to know many columns are contributing any information or value to the model and dropped them for further analysis.

3. Dummy Variables for categorical variables:
   Created dummy variables for all categorical features and made several columns based on the values in the main variables. Dropped the main variables. By this we converted categorical variables in to numerical variables.

4. Data split in to train and test:
   Given data split in to train and test data. We had split 70% data in to train and 30% data in to test data. Made converted column as y variable. Other numeric column as x variables. Used train_test_split function to split the trains and test data.

5. Scaling the numerical variables:
   We scaled there numeric variables “TotalVisits”, “Total time spent on website”,” Page views per visit”. Used the function “StandardScaler function to scale the variables. By scaling, data converted between 1 to 0.

6. Correlation Matrix: Since we created any dummy variables, we have too many values for correlation. Correlation matrix created using heat map is very difficult to read.

7. RFE for feature selection:
   Recursive feature elimination used to select the main features which contribute more for the model. Used RFE functions under feature selection library. Ran RFE with 20 variables.

8. Model Building:
   Used StasModels to access the model. Identified the pvalues of columns which are very high. Dropped those columns with pvalue more than 0.05 and repeated the steps again to select the model with right features.

9. Model Evaluation:
   Calculated the accuracy of the model using metrics, accuracy_score. Created the confusion matrix. Calculated TP, TN, FP, FN. Calculated the Sensitivity, Specificity, False Positive rate, Positive Predictive values, Negative Predictive value. Calculated Precision and Recall.

10. Predictions on test data:
    Scaled the test data, made predictions on the test data. Identified the overall accuracy, specificity and sensitivity of test data.

11. Observations found:
    Identified accuracy as 80% , Sensitivity as 80% and Specificity as 81%

Recommendations:

Identified following features which are contributing more for the conversion rate.

Lead Source_Welingak Website
Lead Source_Reference
What is your current occupation_Working Professional
Last Activity_Other_Activity
Last Activity_SMS Sent
Total Time Spent on Website
Lead Source_Olark Chat
const
Last Notable Activity_Modified
Last Activity_Olark Chat Conversation
Lead Origin_Landing Page Submission
Specialization_Others
Do Not Email

Company should make efforts to convert following top leads:

Lead Source_Welingak Website
Lead Source_Reference
What is your current occupation_Working Professional
Last Activity_Other_Activity
Last Activity_SMS Sent