LEAD SCORING CASE STUDY- SUMMARY REPORT

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site, and the conversion rate.

The following approach was used:

The analysis is performed on the past data (~9000 customers).

- Understanding of data attributes with help of data a dictionary
- Data inspection, cleaning, NULL/Unknown treatment, outlier treatment and etc
- Exploratory data analysis to derive useful insights
- Data Preparation, dummy data creation from categorical variables
- Model creation using logistic regression
    - Training 70% of data (fit and transform) and Test (transform) on 30% of data
    - Model evaluation.
- Model evaluation, predictions, and measure effectiveness

1. **Cleaning data:**

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. A few of the null values were changed to respective median/mode values, so as not to lose much data. Although they were later removed while making dummies.

2. EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values contain outliers so we caped outliers with the respective value.

3. **Dummy Variables:**

The dummy variables were created and later Dropped the repeated variables, for which dummies were created and binary features mapped into binary values.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively. For numeric values, we used the StandardScaler.

5. **Model Building:**

Firstly, RFE was done to attain the top 20 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).

6. **Model Evaluation:**

- A confusion matrix was made. Later on the optimum cut-off value (using the ROC curve) was used to find the accuracy, sensitivity, and specificity which came to be around 80% each.
- The prediction was done on the test data frame with an optimum cut-off of 0.35 with accuracy, sensitivity, and specificity of 80%.
- Precision & Recall was also used to recheck and a cut-off of 0.42 was found with a Precision of 70% and recall of 80% on the test data frame.
- Company can change the cut-off based on business strategy. When we need to get aggressive in business then, we can rely on high sensitivity. And when the company is in a conservative approach and only spends efforts in case to reduce non-conversion then they will rely on high precision.
-

Conclusion

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

- 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

Keeping these in mind 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.