Lead Scoring case study – Summary

This is the summary of the leads scoring case study.

X education has come to us with a problem statement of having a low conversion rate of 30%.

We have studied and worked on the raw data provided by X education to find the top features that affect the conversion rate. Identifying these features will help X education work towards their goal of reaching 80% conversion rate.

X education can use these features to nurture potential leads well in order to get a higher lead conversion.

We have built a logistic regression model to identify the top features that contribute to the conversion rate.

Please find the steps to building this regression model:

1. **Data cleaning:**
    - **a.** Replace the missing values in ‘Lead Quality’, ‘Country’, ‘Specialization’, 'How did you hear about X Education', 'Tags', 'What is your current occupation' as ‘unknown’
    - **b.** Replaced missing values in ‘City’, ‘Lead Profile’, 'What matters most to you in choosing a course' as ‘Other’
    - **c.** Replacing ‘'Asymmetrique Activity Index' values with numerical values(1,2,3 for High, Medium and Low values)
    - d. Dropped the columns 'Prospect ID', 'Lead Number', 'Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score' - as they did not all value
    - e. Dropped rows of columns with less than 2% missing values
    - f. Created a list of all categorical columns and stored it in object_columns - used the LabelEncoder preprocessor to replace all feature values with dummy variables
2. **EDA:**
    - a. EDA on Numerical variables
    - b. EDA on Categorical variables
    - c. Correlation matrix to check classification and correlation
3. **Splitting the data**
    - a. Train_test_split
    - b. Trainsize: 70, Testsize:30
    - c. Scaled the variables using the MinMaxScaler
    - d. X_train = All features, y_train = Target variable
4. **Model building**
    - a. Course tuning using RFE
    - b. Fine tuning using P-values and VIF
5. **Model Evaluation**
    - a. Confusion matrix
    - b. Checking the overall accuracy, sensitivity and specificity

6. **Optimal Cutoff**
    a. ROC function
    b. Creating columns with different probability cutoffs
    c. Plotting the different probability cutoffs and their accuracy, sensitivity and specificity
    d. With the cutoff as 0.35 we have accuracy, sensitivity and specificity of around 76%.
7. **Prediction on Test dataset**
    a. Scaling the columns as per the train data using MinMaxScaler
    b. Splitting the dataset into X and Y datasets
    c. Substituting all the columns in the final train model (adding back the columns that were previously filtered out by RFE)
    d. With the cutoff at 0.35, we have the accuracy at 75%, sensitivity at 80% and specificity at 72%
8. **Precision Recall**
    a. Precision – Recall Tradeoff
9. **Prediction on Test**
    a. Prediction on test set using the updated cutoff
10. **Find the top features**
    a. Created a dataframe and plotted it to find the top features contributing to the conversion rate

Conclusion

Below are the top features that contribute to the conversion rate:

- Lead Origin
- Lead Source
- Do Not Email
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- What matters most to you in choosing a course
- Last Notable Activity

X education may focus more on these top features and implement the below mentioned strategies to possibly improve the conversion rate

Strategy for Aggressive Lead Conversion

- Lower the Cutoff Threshold
- Prioritize by Lead Score
- Monitor and Adjust response and conversion rates

Strategy to Minimize Unnecessary Phone Calls

- Set a High Cutoff Threshold
- Focus on High Lead Scores