Lead Scoring Case Study - SUMMARY

Objective:

This Case Study was focused on solving a strategic problem for the organization X Education. This company tries to attract more customers pursue the online courses offered by them through various promotional channels like online advertisements, chats, promotions etc. Based on the activities of people who view these ads, the company approaches the customers who they sense potential targets for conversion into actual customers who would buy their course and pursue further.

The key objective of this Case Study was to identify some key factors that would help us maximize the success rate and classify the Hot Leads accurately so that the company's conversion rate goes high.

Solution Approach:

Data Cleaning:

- To address this problem, we carried out a detailed analysis on the data shared with us. We were provided with around 9000 data points.
- A thorough data cleaning was done as there were a lot of attributes that were null or had missing information and hence were not useful in drawing any inferences for solving the problem. Such attributes were dropped in the initial data analysis phase.
- Some variables having only one unique value were also dropped as they did not add any variance to the data and hence no value for model building.
- Small number of missing values were imputed with either mean or the most popular value for that variable. Outliers were not dropped because of the nature of changing data in this Business context.
- Dummy variables were introduced for categorical data and all variables were converted to numeric form for model building.

Model Building:

- This problem was solved using the Logistic Regression and RFE (Recursive Feature Elimination) techniques.
- We started off with an initial set of 15 features to build the model.
- The first model gave a good accuracy score but it did show some trace of a high p-value and multi-collinearity. So a variable (lead_source_reference) was dropped and a new model was built using 14 variables.
- The second model (with 14 variables) had the same accuracy score as Model 1. It also had in range p-values for all variables and VIF under 2. So model 2 was taken as the final model for further tests.

Observations:

- The ROC curve showed strong indication of model aligning to the top left corner indicating the model's capability of giving high true positives.
- With a trade-off curve between Accuracy, Sensitivity and Specificity, we can a threshold (cut – off) of 0.35 which was used to classify a lead a Hot lead (conversion probability > 0.35 → Hot lead, else not).

Conclusions:

With the final model we were able to generate the Lead scores for data points available (Lead Score = conversion probability * 100).

We had a cut-off probability of 0.35 as the optimal. So any Prospect with a Lead score of 35 or more is a Hot Lead and should be approached for conversion.

The top variables that we saw driving the results were:

- ✓ Lead origin
- ✓ Lead Source
- ✓ Last Activity

Submission By:

Barkha Garg

Suryateja Rallapalli