LEAD SCORING CASE STUDY SUMMARY

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals interested in the courses land on their websites and browse for courses.

The company sells online courses to several websites and search engines like Google, to industry professionals. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified as a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

X Education needs our help to select the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

From above problem description, we can conclude that this is a problem based on classification. To handle such problems, we can use logistic regression to analyze the Lead Score with help of conversion rate.

The steps used to resolve problem:

1) DATA READING AND UNDERSTANDING:

- Total number of Columns and Rows
- Data Types of columns

2) DATA CLEANING:

i) Checking for any column

ii) Checking for null values

3) DATA VISUALIZATION AND OUTLIERS' TREATMENT:

- We performed univariate analysis, bivariate analysis, and correlation on categorical column.
- Modification of 'select' in columns.

4) FEATURE SCALING:

- Removed columns with more than
- Created dummies using PD.get_dummies function.

5) MODEL BUILDING:

- Used techniques like RFE to check the stability of the model using stats libraries, created dummy variables, and reviewed p-values below 0.05 and vif values to be under 5.

6) MODEL EVALUATION ON TRAIN SET

- Finding Optimal Cut off Point
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut-off is at 0.35.

7) PREDICTIONS ON TEST DATA SET:

TRAIN Data

- Accuracy – 78%
- Sensitivity – 77%
- Specificity – 78%

8) FINAL OBSERVATIONS:

We can get an understanding between a hot lead and a cold lead, the total time spends on the Website, the Total number of visits,

When the lead source was from the below sites:

a. Google

b. Direct traffic

c. Organic search

d. Welingak website

- When the lead origin is Lead add the format, current occupation is as a working professional.