Lead Scoring Case Study Summary

Problem Statement:

- X Education sells online courses to industry professionals. The basic data provided gave us a lot of information about X education business situations like how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.
- We can see X Education struggles with a low lead conversion rate despite significant website traffic.
- So, the company aims to
    - **Increase the percentage of leads** that convert into paying customers from the current rate of around 30% to approximately 80%.
    - **Identify 'Hot Leads'** with higher conversion potential. This involves building a lead scoring model to prioritize leads based on their conversion likelihood.
    - To make the lead conversion process more efficient, the company intends to **focus more on communicating with potential leads** identified as 'Hot Leads' rather than making calls to everyone.
    - **Deploy a model** for future potential lead identification.

Thus, the Lead scoring has been done using logistic regression model to meet the constraints as per business requirements.

The following are the steps used for model building:

1. **Data Cleaning:**
   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. Dropped columns which has more than 40% data missing.
2. **Categorical Analysis:**
    - It was found that a lot of elements in the categorical variables were irrelevant. The outliers are treated using formula Q3+1.5IQR & Q1-1.5IQR.
    - Maximum number of leads are generated by Google and Direct traffic.
    - Conversion Rate of reference leads and leads through welingak website is high.
    - To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
    - Working Professionals going for the course have high chances of joining it. Unemployed leads are the most in terms of Absolute numbers.
    - Leads converted of last notable activity to whom the SMS was sent.

- API and Landing Page Submission bring higher number of leads as well as conversion.
- The most numbers of leads are from India and in terms of city highest number are from Mumbai.

3. Train-Test split and function creation:

Train and test split were done. We have taken 70% data for training and 30% data for test.

4. Model Building:

Taken the logistic regression model and RFE from sklearn to select top 30 features. Then dropped features with high VIF and then dropped features with p-value > 0.05.

5. Model Evaluation:

Matrix used to evaluate model:

- Confusion matrics
- Accuracy score
- Precision
- Sensitivity/Recall
- Specificity

6. Prediction:

Prediction was done on the test data frame with an optimum cut off as 0.42. The sensitivity has increased from train set to test set.

Observations on the train set: Accuracy: 87.56%, Sensitivity: **84.54%,** Specificity: 89.43%

Observations on the test set: Accuracy: 87.91%, Sensitivity: **86.02%,** Specificity: 89.14%

It was found that below variables can be used to find **Hot Leads**, the potential buyers:

'Total Time Spent on Website', 'Lead Origin_Lead Add Form',

'Lead Source_Welingak Website', 'Do Not Email_Yes',

'Last Activity_Converted to Lead', 'Last Activity_Email Bounced',

'Last Activity_Olark Chat Conversation',

'What is your current occupation_Working Professional',

'Tags_Closed by Horizzon', 'Tags_Graduation in progress',

'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_invalid number',

'Tags_switched off', 'Last Notable Activity_Email Bounced',

'Last Notable Activity_Had a Phone Conversation',

'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unreachable'