LEAD SCORING CASE STUDY SUMMARY

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

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

The Goal was to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

SUMMARY:

Step 1: We read and understood the Data

Step 2: Data Cleaning

1. We dropped variables with unique values
2. Many of the categorical variables had a level called 'Select' which is as good as a null value. We Replaced 'Select' with NaN
3. Dropped variables with missing values more than 35> (following the industry standard)
4. Dropped variables that had skewed data and imputed the missing value of the other variables.
5. Merged low frequency labels into one and removed other kinds of data redundancies

Step 3: Preparing the Data for Modelling

1. Converted binary variables to ‘0’ and ‘1’
2. Created Dummy Variables of all categorical variables

Step 4: Splitting the Data

Splitted the data into Train and Test set in the ratio of 7:3 respectively.

Step 5: Feature Scaling

1. We used Standard Scaler to scale the numerical values
2. We mapped a heatmap to check correlation between the variables and dropped the ones with high correlation

Step 6: Model Building

1. We used RFE and selected the 15 top features.
2. We looked at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
3. 7 variables proved promising. The VIF's for these variables were also found to be good.

4. We checked the optimal cut off by finding points and checking the accuracy, sensitivity and specificity and got our final model
5. We plotted the ROC curve for the features and the curve came out to be pretty decent with an area coverage of 84%. Got the cut off value of 0.3
6. Assigned the lead score between 0-100
7. We tested on the test data set and got accuracy value Accuracy : 69.11%, Sensitivity :81.64%, Specificity : 60.94.13%, Precision : 57.71%, Recall :81.64%.
8. Best of all, our model predicted **83%** conversion rate on the test set. Exceeding our target.

Step 8: Inference:

The conversion rate on our final predicted model (83%) meets the goal of 80% conversion rate.

Recommendation would be to focus on the following lead points for **optimum conversion**:

- What is your current occupation_Working Professional
- Total Time Spent on Website
- Lead Source_Reference
- Lead Source_Welingak Website