LEAD SCORE CASE STUDY SUMMARY

This analysis is performed for X Education to track down ways of getting more industry Professionals to join their courses. The fundamental information gave provided us with a great deal of data about how the potential clients visit the site, the time they spend there, how they arrived at the site, and the transformation rate.

Following are the steps performed during the analysis:

1. Read and understand the data

- Imported the data set and performed required methods to understand the data like checking the shape, data types, statistical information, data dictionary of the lead score data

2. Data Cleaning and Data Manipulation

- Check and handle the duplicate data, missing values and outliers in the data
- Since there were significant values with select implying the data was missing, so we had to replace them with null values for further analysis of the data
- Dropping the columns 'Magazine', 'Receive More Updates About Our Courses', 'I agree to pay the amount through cheque', 'Get updates on DM Content', 'Update me on Supply Chain Content' as they all contain unique value and do not require for analysis.
- Drop the columns prospect ID and Lead Number as they are unique IDs which are not necessary for analysis.
- The columns which are having more than 35% missing values have been dropped
- For all the columns which had less than 25% of missing values however were not dropped because of their relevance for analysis had been replaced with a new category 'missing '

3. Exploratory Data Analysis

- Univariate Analysis and Bivariate Analysis have been performed and observations were recorded in the python file.

4. Create Dummy variables

- Dummy variables are created for object variables.

5. Train -Test Split

- The split was done at 70% and 30% for train and test data respectively

- Performed Feature Scaling technique using Minmax Scaler

6. Logistic Regression Classification technique for the model building

- Using RFE 15 relevant variables were attained. Later the rest of the variables were removed manually depending on the p-value and VIF values (The variables with p-value < 0.05 and VIF < 5 were kept).

7. Model Evaluation

- A confusion matrix was made followed by obtaining the optimum cut off value (using ROC curve) the accuracy, sensitivity and specificity were determined and those came to be around 80%

8. Prediction

- Prediction was done on the test data frame and with an optimum cut off as 0.37 with accuracy, sensitivity and specificity of 80%.

9. Precision – Recall

- This method was also used to recheck and a cut off of 0.41 was found with Precision around 71% and recall around 77% on the test data frame.

Conclusion:

It was observed that the factors that made the biggest difference in the potential purchasers are -

1. Total time spent on the website
2. When the lead origin is lead add form
3. When the lead source was -
    - Direct traffic
    - Welingak website
4. When the customer opted for not to email
5. When the customer current occupation is working professional.
6. When the customer says that he is choosing this course for better career prospects.
7. When the last activity was -
    - Olark Chat conversation
    - Phone conversation
8. When the last notable activity was modified, email opened, page visited on website, email link clicked and olark chat conversation.

Keeping this in mind, the X Education can thrive as they have an exceptionally high opportunity to get almost every one of the likely purchasers to adjust their perspective and purchase their courses.