SUMMARY

Company name : X Education

The analysis seeks to support X Education in boosting the enrolment of industry professionals in their courses. The initial dataset provides key insights into visitor behaviour, such as website interactions, session lengths, referral sources, and conversion rates.

The approach involved the following stages:

1.Data Preprocessing:

It includes details for comprehending the data and its structure using methods like .info(), .describe(), and .head(). This analysis revealed that the dataset contains 37 columns, some of which have null values. We determined that logistic regression would be suitable due to the presence of many categorical variables, necessitating the creation of dummy variables. Additionally, with no outliers present, we proceeded directly to addressing the null values.

2. Data Cleaning:

Handling Null values: The dataset was generally clean, with only a few missing values. A small number of missing records were deleted as their removal would not impact the analysis significantly. To manage null columns, we used a 30% cutoff method, as only five columns had more than 30% missing values, and these columns provided minimal useful information, so we decided to drop them. Additionally, we removed columns with a high proportion of 'select' values, which were treated as null values. Three columns with a significant amount of 'select' values were dropped. The 'city' column primarily contained data from Mumbai, with a high number of "select" values indicating that most entries were from this city.

data imputation: Additionally, we employed data imputation rather than dropping columns due to the small size of the dataset. For the remaining columns with null values, we set a 15% cutoff and imputed these missing values with 'Unknown'.

3. Data Exploration(EDA):

Analysis: EDA was conducted for feature analysis, leading to the decision to drop columns with majority 'NO' responses, as they did not provide meaningful information. It was also observed that many elements in the categorical variables were irrelevant. The numeric values appeared satisfactory, with no outliers detected. At this stage, we discovered that the majority of students were from India, leading to the decision to drop the country column due to its limited informational value. As a result, 98% of the rows in the dataset were retained for further analysis.

4. Data Preparation:

Dummy Variables: Dummy variables were created for the categorical variables, We did not drop the specialization column earlier and handled the 'select' data by imputing 'Unknown' values. Careful attention was given to creating dummies for these cases and dropping the relevant columns. For scaling numeric values, we used **MinMaxScaler**.

Train-Test split: The split was done at 70% and 30% for train and test data, respectively.

5. Model Building & Feature Selection:

Model Build: Initially, we build a logistic regression model

Feature Engineering: Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables. Subsequently, the remaining variables were manually reviewed and removed based on their

VIF values and p-values. Variables with a VIF less than 5 and a p-value less than 0.05 were retained, while features that violated these criteria were dropped.

6. Model Evaluation:

Confusion Matrix: We generated a confusion matrix to assess the model's performance, using a random cutoff of 0.5, which resulted in an accuracy of 80%, sensitivity of 65%, and specificity of 88%. which is imbalance in nature

Optimal Cutoff: By analysing the ROC curve, we identified the optimal threshold as 0.39. This threshold provided an accuracy, sensitivity, and specificity of approximately 80%, effectively balancing all three metrics.

Precision and Recall Analysis: For additional validation, we used the precision-recall approach. With the 0.39 threshold, we achieved a precision of around 73% and a recall of approximately 80% on the training dataset.

7. **Model Evaluation on Test dataset:**

Confusion Matrix: On the unseen data, we created a confusion matrix and evaluated all metrics using the cutoff value of 0.39 obtained from the training data.

Optimal Cutoff: This threshold yielded an accuracy, sensitivity, and specificity of approximately 80%, indicating that the model is neither overfitting nor underfitting and is performing well in generalizing to unseen data.

Precision and Recall Analysis: Additionally, we observed a minor difference in precision and recall, with precision around 71% and recall around 78%. This suggests that the model's performance is stable across different datasets.

To enhance conversion rates, implement the following strategies:

• Focus on leads originating from `Reference` and `Welingak Website` and `Organic search` sources, as they have a higher likelihood of conversion.

• Prioritize leads who are `working professionals`, as they show a higher tendency to convert into customers..

• Reach out to leads who `spent more time on the website` or `page visited` as their engagement suggests a higher chance of conversion.

Give special attention to leads from the `Olark Chat` source, as they exhibit a higher probability of conversion.

• Reach out to leads whose last activity was marked as `SMS Sent` as this activity indicates a better chance of conversion.

• Consider not contacting leads whose last activity was `Olark Chat Conversation` as these interactions are less likely to result in conversion.

• Be cautious with leads from `Landing Page Submission` as they tend to have lower conversion rates.

• Focus on leads with specializations like `Working professionals` and `Unemployed`, as they are more likely to seek better career opportunities and show a higher likelihood of conversion. Be selective with leads labelled as `Others`, `Businessman` or `Housewife` as these categories demonstrate a lower likelihood of conversion.

• Minimize outreach to leads who have opted for `Do Not Email` since they are less likely to convert.

• Also, Improve the design and content of your `lead capture forms` called `Lead form` to enhance data quality and lead scoring accuracy.