SUMMARY

The aim of this evaluation conducted for X Education is to discover methods that can be utilized to draw in additional businesses from the industry. The information given provided significant insight into the enrollment of professionals in courses. The manners and traits of potential clientele, encompassing their web browsing activity. The metrics that are commonly used for website performance measurement include patterns of user behavior, time spent on the site, referral source, and the conversion rate of website visitors.

The following steps are used:

Inspecting the data:

The aim of reading and examining the data is to have a clearer perception of the data and recognize any issues or trends that may require being tackled before carrying on with the analysis.

Data Cleaning:

1. No duplicated rows were found in the data set during the data cleaning process.
2. Adding null values to a dataset can corrupt its information. Seven columns in the given dataset had more than 45% of null values, obscuring its data.
3. Categorical columns having null values less than 45% and outlier in the dataset are treated by replacing them with the most frequent values within each column. Numerical columns having outliers are capped at upper limit 0.99.
4. The following columns were removed from the data set as they were set up to have lesser than 45 null values' Prospect ID',' Lead Number'(as they were only identification figures and handed no information for the model),' How did you hear about X Education',' Lead Profile',' Lead Quality',' Asymmetrique exertion indicator',' Asymmetrique Profile Index', ' Asymmetrique exertion Score',' Asymmetrique Profile Score'.

EDA:

1. The data set has a slight imbalance, with 3346 rows indicating a conversion (1) and 5487 rows indicating non-conversion (0).

2. The analysis found that working professionals have a higher likelihood of converting into a "hot" lead, while unemployed individuals have a higher likelihood of not converting into a "cold" lead.

3. It was observed that a majority of "hot" leads had a specialization related to management.

4. The Lead Add Form had a higher conversion rate, while the API had a lower conversion rate or fewer "hot" leads.

5. The majority of leads came from Google.

6. The "Reference" and "Welingak Website" as well as other sources showed higher conversion rates.

7. The majority of "cold" leads came from Olark Chat, Organic Search, Direct traffic and Google.

Useless Variables:

Dummy variables are created to convert categorical variables into numerical variables that can be used as features in machine learning models. This allowed the model to treat categorical variables as numerical input and use it in the analysis

Train-Test Split:

The data was split into 80% for train and 20% for test data. The goal of train-test split is to assess how well the model will perform on unseen data, by evaluating its performance on the test set. This helped to ensure that the model is not overfitting the training data, and will generalize well to new, unseen data.

Model Building:

Model-2 is considered as final model as the p-values of all independent variables are almost zero and no high VIF variables or multi-collinearity

Model Evaluation:

A confusion matrix was created to evaluate the model. The optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to have optimal values of 0.9141, 0.9011, and 0.9170 respectively.

Prediction:

Prediction was done on the test data-frame with an optimum cut-off value of 0.3. With this cut-off, the evaluation metrics 'accuracy', 'sensitivity', and 'specificity' have optimal values of 0.9141, 0.9011, and 0.9170 respectively. This means that the model can correctly predict the target variable with a high level of accuracy and has a high level of sensitivity and specificity

Conclusion:

It was found that the variables that mattered the most in the potential buyers are (in descending order):

The total time spent on the Website. Total number of visits. Lead source: Google, Direct traffic, Organic search, Welingak website. Lead origin: Lead add format. Current occupation: Working professional.

The model's results demonstrate its strong performance on both the train and test data. The accuracy, sensitivity, and specificity of the train data were 0.9141, 0.9011, and 0.9170, respectively. The test data's corresponding metrics were 0.9241, 0.8621, and 0.9621. These outcomes indicate that the model can accurately identify hot leads, resulting in a positive impact on the company's efforts to prioritize the most promising leads and, ultimately, increase their chances of closing a sale.