CASE STUDY SUMMARY REPORT

Problem Statement: Identification of the “hot” leads and increasing the leads conversion ratio from 30 % to 80 % for an online education provider company “X Education”.

Approach to the problem:

We have used the Logistic Regression Machine-Learning algorithm to create model with higher conversion rate followed by recommendations and analysis.

The overall approach followed was as below:

The pre-step involves importing all the required Python libraries, functions for the case study for data visualization and machine learning model building etc.

1. **Data Preparation**: The Data Understanding, Data Preparation and Exploratory Data Analysis (EDA) is done in this step.
Here first the data is loaded and dataset is examined for the shape, statistical information, data-types, and columns etc.
Missing value analysis in terms of percentage and number for each column was done.
We have created missing_data function to make **code modular**.

Data quality issue were addressed like “select” was replaced with nan.
Then other data quality issue was addressed like binary values converted to 0’s and 1’s as ML model supports only numerical attributes.
This was followed by the **Univariate and Bivariate analysis** for each variable with **visualization** for each variable to perform **missing value imputations**, data redundancy handling etc.
The data is converted to a clean format suitable for analysis in Python. **Outlier analysis** for numeric variable was done using boxplot followed by outlier treatment.
Dummy variable for some of the categorical variables was done for categorical variables using one hot encoding.

2. **Dividing the data into test and training data** for our model:
Using train_test_split the data was divided in the 70-30 % ration for the train and test data.

3. **Feature Scaling** and **Normalization of Data**:
StandardScaler was used for scaling the numeric features in the data set.

4. **Model building** using machine learning algorithm:
Here the first step is **automatic feature selection** using RFE to get top few features for our model.

The Iterative process was followed to build the final **Logistic Regression model** considering minimising the multi-collinearity (VIF Score) and p-value.
The final modal gave various features which were significant for our model with respective weightage for each of them:

	coef
const	-3.4708
Last Activity_Converted to Lead	-1.4452
Last Activity_Email Bounced	-2.6793
Last Activity_Olark Chat Conversation	-1.8792
What is your current occupation_Working Professional	1.5701
Tags_Busy	2.7424
Tags_Closed by Horizzon	8.1946
Tags_Lost to EINS	10.1310
Tags_Ringing	-1.1238
Tags_Will revert after reading the email	5.3471
Tags_switched off	-1.7187
Last Notable Activity_SMS Sent	2.0440
ıat matters most to you in choosing a course_Other_Factors	-3.9167

5. **Model Evaluation** based on various metrics:
Receiver Operating Curve (ROC) was created to determine the accuracy of the model and Area Under Curve analysis. Then plot between probabilities, sensitivity, specificity was done to find the optimal point to take it as a cutoff probability. Next the performance of the test data was checked based on the model built and test and train performance was compared for various evaluation metrics like accuracy, sensitivity, specificity, recall and precision.

6. **Recommendations**, **Lead score assignment** for the current data:

Finally lead score was assigned by multiplying the conversion probability with 100 for each data point. We can conclude that this is good model as the performance on the test data is more or less better than the performance on the training data-set.