Summary Report

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. We need to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The following are the steps performed:

1. **Cleaning data:**

The data was partially clean except for a few null values. There were plenty of 'Select' values in the data which had to be replaced with null values since it did not give us much information.

Columns which had more than 40% missing values were removed. ['Asymmetriqu e Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'As ymmetrique Profile Score', 'Lead Profile', 'Lead Quality', 'How did you hear about X Education']

After removing the columns with more than 40% missing values, few columns still had large number of missing values. For those, we went through the columns one by one, found out the maximum value for that specific column and imputed the null values with that value. Some null values were changed to 'Others' for better understanding of the data.

There are few sales team generated columns present in the data.
When we will be running the ML model to understand which candidate the sales team should call first and which candidate should they call last, the data that we will be using will be directly from the form. The data would not be having the sales team generated data. Hence, we need to remove all the sales team generated data before building our model.
Sales team generated columns are: ['Tags', 'Lead Quality', 'Last Activity', 'Last Notable Activity', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score' and 'Asymmetrique Profile Score'].

2. Exploratory Data Analysis:

A quick EDA was done to check the condition of our data.
It was found that numerical variables had outliers. We handled outliers by removing the top and bottom 1% of the data.
It was also found that a lot of categorical variables had only a particular value repeated the greatest number of time and thus were irrelevant for our analysis. We removed such variables from the data before building our model.

3. Data Preparation and creation of Dummy Variables:

Binary variables were imputed with '1/0' values. For categorical variables dummy variables were created.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively. Numerical variables were rescaled in the train data frame using Standard Scaler.

5. Model Building:

We used RFE to identify top 15 relevant variables.
We built the logistic regression model using the variables selected by RFE. We removed some variables manually depending on the p and VIF values. (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation:

A confusion matrix was created, and the accuracy, sensitivity and specificity of the model was found out based on '0.5' threshold.
Later the optimum cut off value was found using ROC curve which came out to be '0.3' and using that value the accuracy, sensitivity and specificity was recalculated. The respective values were: Accuracy: 79.66%, Sensitivity: 76.50% and Specificity: 81.60% for the train data frame.

7. Prediction:

Prediction was made on the test data frame using the model build. The optimum cut off '0.3' was chosen which resulted in accuracy, sensitivity, and specificity values as 79.08%, 76.55% and 80.58%.

Below can be considered as hot leads by the sales team:

- Customers whose source of lead was 'Welingak Website' have the highest possibility of converting.
- Customers whose source of lead was 'Reference' also have the highest possibility of converting.
- Customers spending major time on website should be targeted.
- Customers whose source of lead was 'Olark Chat' also have high probability of converting.