**Summary Report**

The Analysis and Logistic Regression Model is built for the **X Education sells online courses** companyto help the company select the most promising leads. We have created the model using features provided in the dataset and assigned **Lead Score** to every lead according to the probability of them getting converted.

The following steps were used in Model Building –

1. **Data Cleaning:**

Replaced the value ‘Select’ with null as those are missing

Dropped the columns which have more than 45% of missing values in them

Dropped the categorical columns with highly skewed data

Imputed or dropped the data where there were less percentage of missing values

Dropped the columns provided by the Sales teams after leads were converted, as they are not needed for the model building process

1. **EDA:**

We have done Univariate, Bivariate analysis on the categorical and numerical features to find the impact of the variables to the target variable.

Also, performed outlier treatment on the numerical features

1. **Data Preparation:**

Created Dummy Variables for the multiple level categorical features

1. **Train-Test Split:**

We split the remaining data at 70% & 30% for train & test data set respectively

1. **Model Building:**

Built the model on the Train set

Performed RFE with 15 features

Then the features were manually removed depending on the p-value (greater than 0) and VIF (greater than 5) of the features

1. **Model Evaluation:**

Evaluated the ROC curve to find the optimum cut-off value which came out to be 0.36

Created Confusion matrix and calculated the performance metrics Accuracy, Sensitivity & Specificity

After the final model creation, the value of Accuracy, Sensitivity & Specificity were around 80%

1. **Prediction:**

Prediction was done on the test set using the cut-off value of 0.36 and the performance metrics came out to be around 80% similar to train set

Also, added Lead Score column to the actual dataset which is calculated by the predicted probability \* 100

We could see that the below features were the most impactful in Lead Conversion -

1. When Lead Origin is Lead Add Form
2. When Lead Source is
3. Google
4. Direct Traffic
5. Reference
6. Welingak Website
7. When Last Activity or Last Notable Activity is SMS Sent
8. When Current Occupation is Working Professional
9. Total Time Spent on Website

So, the Company can use the Lead Score provided by the model to determine which lead to be approached on priority and focus on the people from above features to increase the lead conversion rate of the company.