

## **Summary**

We did analysis from the given dataset and built a model to understand the factors that can increase the lead conversion rate of courses from an EdTech company named X Education.

*Steps:*

### **Cleaning data:**

The data was mostly clean, we took care of a few elements like replacing the null values with our constant 'not available', and shedding off the columns which had a significant amount of null values that could have interfered with our analysis.

### **EDA:**

Exploratory Data Analysis of the cleaned dataset was used to see how numerical and categorical variables are distributed. On checking, there were no significant outliers hence we proceeded to the next step.

### **Dummy Variables:**

The variables of 'object' datatype was chosen and dummy variables were created for them for further analysis. MinMaxScaler was used for numeric variables.

### **Train-Test split & model building:**

The train and test data was split into 75% and 25% respectively. Recursive Feature Elimination was done to attain the top 15 relevant variables. A few more variables were removed when statistically it crossed the threshold limit of p-values and vif.

## **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.5 with accuracy of 80%.

*The variables that directly impact the lead conversion (Ordered in descending order) :*

- >The total time spent on website.
- >Total number of visits.
- >When the lead source was: Direct traffic, Welingak website,
- >Organic search
- >When the last activity was: Olark chat conversation
- >When the lead origin is Lead add format.

Prioritizing these factors would lead to a high lead conversion rate.