Summary

Problem statement:

X Education sells online courses to industry professionals. X Education gets a lot of leads, its lead conversion rate is poor. To make this process more efficient, the company want to identify the most potential leads. They want to build a Model which identifies the hot leads. Deployment of the model for the future use. If they successfully identify this set of leads by the model, 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.

**The following are the steps used:**

1. Cleaning data
2. EDA
3. Train-Test split
4. Model Building
5. Model Evaluation
6. Prediction
7. Precision – Recall

**Clearing data**:- In this delete some columns which contains the null or missing values more than 40 % and columns which contains only unique values also replacing the select value in data set with ‘Nan’ or ‘data not given’ values.

**EDA**: - In this step checking missing values of columns and do some univariate analysis and bivariate analyse of features with target variable converted.

**Train-Test split: -** Split whole data into train and test data set using scikit-learn library of python. Taken 70% train data and 30% test data. And also create dummy variables of categorical variables have values more than 2. Rescale the whole data using MinMaxScaler from sklearn library.

**Model Building: -** Build logistic regression model using sklearn and stats model libraries and test the model in test data set.

**Model Evaluation:** - Evaluate the build model using confusion metrics and taking **cut off 0.5** randomly and find **accuracy = 80.2%, sensitivity = 64.8% and specificity = 89.7%.**  **ROC CURVE (area = 0.88) and** further plotting plot **on x-axis = probability, y-axis= ['accuracy','sensitivity','specificity']** get **optimal cut off = 0.39**. Again recheck and get **accuracy of 81.2 %, sensitivity = 79.5% and specificity = 82.3%.**

Check on test dataset: - with **cut off = 0.39**, get **accuracy = 81.8%,** **%, sensitivity = 79.45% and specificity = 83.3%.**

**Prediction: -**

1.Maximum leads originated from landing page submission and lead source were Google.

2. Most of people who applied for course are unemployed in occupations.

3. Maximum leads converted who landing on page submission.

4. Most of leads in course are looking for better career prospects in future

5. Maximum leads who searched for course are from India.

**Precision – Recall:-**

**Train dataset**: - **Precision =** **75.7%** and **Recall value** = **76.56%.**

**Test dataset**: - **Precision = 78.5%** and **Recall value = 76.8%.**