SUMMARY

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 wants its leads generation rate to be increased. For which they need to identify what are the potential leads or Hot Leads.

For this company wants to create a model where we can assign lead score to each of the identified leads for higher chances of conversion. Target for the company is 80 percent.

Data:

For this we have been provided with a leads dataset from the past with around 9000 data points.

This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last activity etc.

Goals of the case study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps to follow to build a Model

1. Reading and Understanding the data
2. Data Cleaning
3. Data Visualization using EDA
4. Data Preparation for Modelling
5. Model Building
6. Model Evaluation

1. Reading and Understanding the data

- We import all the necessary libraries for e.g. NumPy, pandas, matplotlib seaborn etc., import all thewarnings.
- We read the data and check the no. of rows and columns. We also, check if there are any missing/ nullvalues or not. Afterwards, we see the statistical summary of the data.

2. Data Cleaning

- We saw there were few columns with high percentage of null values, so we decided to drop those columns.

- Few columns had null values but the columns were important for analysis so we replaced all null Values with ‘Not Provided’.
- Few of the columns had values as ‘Select’ so we replaced it by ‘NaN’
- Few columns were having outliers and the treatment of outliers was perform

3. Data Visualization using EDA

- For data visualization we performed exploratory data analysis (EDA). We first did the univariate analysisof categorical and continuous data.
- Quick check was done on % of null value and we dropped columns with more than 45% missing values.
- Moved ahead with the bi-variate analysis of categorical and continuous data.
- We also found in country data most of the records were from India and few were from outside India andwe classified as same.
- We found correlations between variables by using different plots.

4. Data Preparation for Modelling

- Data preparation for multiple linear regression involves handling the categorical variables first and thenperforming dummy encoding.
- We then performed the train and test split using 70%-30% rule and then performed the scaling of variables. Since scaling of variables is an important step, we may have different variables of different scales. So, it’s important to have everything on the same scale for the model to be easily interpretable.
- Therefore, we used **MinMaxScaler** for the same.

5. Model Building

- We follow the bottom up approach for this, i.e. we start by building the model with just one variable.Hence, the choice of variables becomes very crucial.
- RFE was done to attain the top 15 relevant variables. First, we will look at the significance of variablesand based on the significance.
- We checked VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). In ourmodel for all the feature VIF < 5 and p-value < 0.05.

6. Model Evaluation

- We first created confusion matrix to find the TP, TN, FP

  and FN.We calculated the sensitivity and specificity.
- We plotted the graph of sensitivity, accuracy and specificity for each level of probability. We found that
- 0.335 was the cut-off point.
- We then used the cut-off point 0.335 to select the person and see if he would be converted or not then,
- We again created the confusion matrix to calculate TP, TN,FP & FN calculated the sensitivity and specificity and got 0.81 and 0.78 respectively.

Sensitivity – Specificity

If we go with Sensitivity- Specificity Evaluation.

- On **Training Data**
  - The optimum cut off value was found using ROC curve. The area under ROC curve was 0.88.
  - After Plotting we found that optimum cutoff was **0.35** which gave

    Accuracy 80.91%
    Sensitivity 79.94%
    Specificity 81.50%.
- Prediction on **Test Data**
  - We get

    Accuracy 80.02%
    Sensitivity 79.23%
    Specificity 80.50%

Precision – Recall:

If we go with Precision – Recall Evaluation

On **Training Data**

- With the cutoff of 0.35 we get the Precision & Recall of 79.29% & 70.22% respectively.
- So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of **0.44** which gave

  Accuracy 81.80%
  Precision 75.71%
  Recall 76.32%

Prediction on **Test Data**

Accuracy 80.57%
Precision 74.87%
Recall 73.26%

So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.35

If we go with Precision – Recall Evaluation the optimal cut off value would be 0.44

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- LEAD SOURCE:
  - Total Visits
  - Total Time Spent on Website
- Lead Origin:
  - Lead Add Form
- Lead source:
  - Direct traffic
  - Google
  - Organic search
  - Referral Sites

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.