Summary

Problem Statement

An education company named X Education sells online courses to industry professionals. Through marketing and the leads procured by various resources, the conversion rate is around 30%, which is considerably poor.

With this case study, X Education wants to identify the most significant leads and focus on them in near future in order to increase the customers opting for the online offered courses by them.

This study potentially should increase the conversion rate as well as help sales team to dial up those whose probability is higher as per the hot leads instead of identifying in random.

Goals of the Study:

- *Identifying the most potential leads or hot leads which contribute more in higher conversion rate*
- *Build a model with the identified potential leads using logistic regression and evaluate the metrics.*
- *Evaluate the model and deploy the algorithm in order to assist sales team improvise conversion rate.*

Steps followed in analysis to achieve the Goals are explained below:

1. *Importing the necessary mathematical, visual and statistical support libraries to perform the analysis*

- *Numpy*
- *Pandas*
- *Matplotlib*
- *Seaborn*
- *Sklearn*
- *Statsmodel*

2. *Reading and understanding the Dataset*

Using Pandas, dataset was imported and basic checks were done using shape, describe and info functions. It was found that data cleaning is highly important step to control the missing data impacting the analysis and providing meaningful insights.

3. *Data Cleaning*

- *It is observed in previous stage of understanding the dataset step that, during the data collection some of the information are left by default with 'Select' as a value which are actually the null values. Let's start the data cleaning process by replacing all the default selections with null values.*

- *Next, we identified quite a good number of null values in dataset with highest being 'How did you hear about X Education' with 78.46%, which could have troubled and highly affected the analysis. Hence, we have dropped the columns with 35% null values straight away.*

- *Specialization, What matters most to you in choosing a course and What is your current occupation, Since they are the most important factors to understand the behavior of a student enrolling into the course, null values are filled as unknown*

- *Upon checking the null value percentage for Total Visits, Page Views Per Visit, Last Activity and Lead Source, it is noticed to be less than 2% and hence dropping entire rows in dataset.*

- *we had then reached the stage of data cleaning with no null values in our dataset loaded for analysis. Moving forward, we checked the unique value counts for each of the categorical variables.*

- *Upon results with variables having unique value count one such as Magazine, I agree to pay the amount through cheque, Get updates on DM Content, Update me on Supply Chain Content, Receive More Updates About Our Courses are dropped from dataset including Country as 70% leads are from India only.*

- *Last but not least, we have dropped Prospect ID and Lead Number as all the values in them are unique and doesn't contribute anything for analysis*

4. *Exploratory Data Analysis*
As soon as data cleaned, it was a perfect time to understand the distribution of data amongst each variable and draw some insights to understand what are the factors influencing the most in conversion rate, which could possibly turn out to be most important while building the model.

<u>Inferences</u>:

- *The dataset is not highly imbalanced and hence meaningful insights can be drawn with respect to target variable.*
- *The major motto which has been mattered to the most in choosing the course is better career prospects.*
- *Landing page submission is a major lead origin to identify the customer to be a lead.*
- *Working professionals and unemployed are highest amongst the population in occupation variable.*
- *Google, organic search along with direct traffic has been a main source of lead for customers.*
- *Most of the customers have enrolled to course after SMS sent or Email being read by them as per the latest activity.*
- *Finance Management is the highest selected specialization category whereas HR and Marketing are amongst top*

5. *Train – Test Split*
Splitting the dataset into 70% train data and 30% test data for building the model using train_test_split function.

6. *Model Building*
Using RFE method of feature elimination and selection of variables. We have considered to reduce it to 15 variables. We have Scaling the numerical variables using the MinMaxScaler function. Building model by adjusting or removing the p-value and VIF by manual elimination method.

7. *Model Evaluation*
A confusion matrix was found out with the final trained dataset and upon calculating the metrics, we had already reached a good value. However, in order to optimize them, we followed a method of ROC curve and optimized the values from 0.5 to 0.35 cut off.

8. *Predictions on Test dataset and finding out the values of evaluation metrices.*
We had followed the same cut off values and calculated the Evaluation results on test set for which the results are as recorded below:

- `Accuracy: 0.79`
- `Sensitivity: 0.79`
- `Specificity: 0.78`

9. *Precision recall evaluation*
With the cut off of 0.5, below values were calculated on test data, which was significant on the model.

- `Precision: 0.72`
- `Recall: 0.71`

<u>Conclusion</u>

Hot leads as per the model are Total Time Spent on Website, Total Visits, Lead Source_direct traffic, Last Activity_sms sent, Lead Source_organic search, What is your current occupation_working professional, Lead Origin_lead add form, Do Not Email_yes, Lead Source_referral sites, Last Activity_had a phone conversation, Last Activity_olark chat conversation and Last Notable Activity_unreachable.

- *Total time spent on website and total visits to website plays a vital role, as more the customer checks the website and spend time, more the chances of enrolling to course.*

- *Source for the leads to look after are Direct traffic, organic search and referral sites, which has higher probability as per model to enrol the customer.*

- *Next, Phone conversation, Olark chat conversation, SMS sent are notable amongst the last activities which can possibly lead to conversion of customer to enrol into course*

- *Last but not the least, Customers who are working professionals are the hot leads for enrolling into the course.*