Project Report

•Hypothesis regarding possible root causes of low lead conversion

1. Not following up with leads who show significant interest with actions like opening an email and sending SMS.

2. Possible inability to guide leads who have identified themselves as unemplyed to forge a new career path on the basis of available courses.

3. Not identifying leads correctly as potential customers based on the metrics in the database.

4. Not giving people who have identified themselves as looking for better career prospects a suitable course and job placement assurances.

- **Map the problem to the relevant data science problem and develop the solution approach that you'll follow.**

Here we have to perform Exploratory Data Analysis on the data set. To find out which of the columns have statistically valuable information that can be used to build a model which can predict with reasonable accuracy given the above mentioned data whether a lead will get converted or not.

To find this out, we can use many machine learning algorithms like Logistic Regression because we need to classify whether a lead will be converted or not.

- **Discuss at least 3 ML models that can be utilised to solve this problem. Note that you don't have to build the model for the dataset, just a brief description of the models that would work in this scenario along with the reasons on when one can be preferred over the other.**

Since its a binary classification problem where the lead can get converted or not there can be many ML algorithms which we can follow which are as follows:

Logistic Regression
Logistic Regression utilises the Sigmoid function to return the probabilty of a category. It generates a probability. By comparing the probability with a pre defined optimum probability, the object is assigned to a category accordingly.

Support Vector Machines
It means to assign a set of hyper planes called decision boundary, that separates the data points into separate classes. The data points closest to the decision boundary are called support vectors. An optimum decision boundary will have maximum distance from the support vectors.

Decision Tree
The algorithm builds branches in a hierarchical manner where each branch is governed by an if else statement. The branched divides the datasets into subsets based on the most important features.

We used Logistic Regression in this instance because :

1. The training data is limited. Logistic Regression works well with less data
2. Features are highly correlated. It handles correlation well.
3. Logistic Regression is adept at handling outliers.
4. The relationship between the predictor variables and the outcome variables can be transformed to a linear form.

- **Proceed with EDA to find the most relevant variables that affect lead conversion.**

Thus EDA was done and the model was built based on the Logistic Regression ML algorithm.
Through the results of the model and the process of RFE we derived the most relevant variables which affect lead conversion.

These are :
Total Visits, Total Time Spent on the website, Lead Origin_Landing Page Submission , Lead Origin_Lead Add Form , Lead Source_Olark Chat , Last Notable Activity_Had a Phone Conversation, Last Activity_SMS Sent What is your current occupation_Unemployed

These variables show the features that help to understand whether a lead will be converted or not in the most statistically significant way.

- **Mention the evaluation metrics that you'll be using to track your model's performance. Make sure that they're mapped to the relevant business outcomes.**

The evaluation metrics that we used to track the models performance are :

Accuracy, Sensitivity, Specificity, Precision and Recall and the AUC-ROC.

The accuracy of our model = 78.45%
Sensitivity = 77%
Specificity = 78.91%
Precision = 80%
Recall = 73%
AUC ROC : 0.86

Altogether they are reasonable evaluation metric values which are indicators of a good model.