SUMMARY

Problem Statement:

X Education is a corporation that sells online courses to industry experts. Despite receiving a high volume of leads, the company's conversion rate is low. The team has been tasked with identifying the most promising prospects who are likely to become paying customers.

Solution:

Preparing and Cleaning Dataset:

- With over 9000 data points, we can delete columns with 30% missing values.
- We removed the City and Country variables as they are irrelevant for our online courses.
- Prospect ID and Lead Number are only record identifiers and have been removed.

- We removed columns with skewed data points as they have no predictive value.
- After cleansing the data, we found a 48% conversion rate.

Exploratory Data Analysis (EDA):

From the univariate analysis we can Hypothesis that

- The majority of leads come via landing page submissions, followed by API calls.
- More leads come from unemployed customers.

From bivariate analysis of the columns with converted column indicates

- Leads from Add Forms are more likely to convert.
- Working Professionals and Housewives are more likely to convert.
- Leads from Live Chat, Reference, WeLearn, and Welingak Websites are more likely to convert.

Model Building:

- We created dummy variables for all categorical variables and we split the data into train andtest sets with a ratio of 70:30
- We scaled the numerical features with **MinMaxScaler**
- W used Recursive feature Elimination (RFE) to identify 15 most important features in thedata set to make the model more robust
- After building our first model we used the Variable inflation factor and p-values of the modelto eliminate the statistically insignificant features
- Finally, we ended up with 11 features for the model.

- We created a lead score (i.e. Conversion probability*100) to give a score between 0 and 100. A higher score indicates a hot lead having a higher probability of lead conversion

Model Evaluation:

- The area under the ROC curve was 86% which indicates this is a good model
- From the sensitivity and specificity tradeoff the optimal cutoff point was 0.44 and the metrics for the train set was

Accuracy	79.09%
Sensitivity	79.34%
Specificity	78.85%
Precision	77.71%
Recall	79.34%

Making Predictions on the Test Set:

- The metrics for predictions on the test set is as follows and they are very close to the training set.

Accuracy	78.95%
Sensitivity	77.71%
Specificity	80.10%
Precision	78.40%
Recall	77.71%

Conclusion:

The primary factor influencing decision-making is the:

1. Total number of visitors.
2. Total time spent on website
3. Lead Origin:Lead Add Form
4. Lead Source: Welingak Website
5. What is the present occupation? Unemployed

Learning:

• Prepare data for logistic regression analysis.
• Develop a Logistic Regression model in Python.

• How to generate dummy variables for category columns.
• How to select a model cut-off based on sensitivity and specificity?
• Obtain a list of variables from the final model that contribute significantly to probability and solve business problems.