SUMMARY

Problem Statement:

X education is a company which sells online course to industry professionals. The company gets a lot of leads but the lead conversion for the company is very poor. They have assigned a team to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Solution:

Preparing and Cleaning Dataset:

- There a lot of columns with high number of missing values and since we have around 9000+ data points we can eliminate the columns with 30% missing values
- We dropped City and Country variables since it's of no use to us as the company provides online courses
- Prospect ID and Lead Number are just records identifierand as hence dropped
- We dropped all columns which have skewed data points as it wont have any predictability value
- We have found 48% conversion rate after cleaning the data

Exploratory Data Analysis (EDA):

From the univariate analysis we can Hypothesis that

- Majority of leads are originated from Landing Page Submission followed by API
- Majority of leads are originated from Landing Page Submission followed by API
- More leads are received from Unemployed customers

From bivariate analysis of the columns with converted column indicates

- Lead originated from Add Form are more likely to be converted
- Working Professional and Housewife are more likely to be converted
- Lead sources from Live Chat, Reference, WeLearn and Welingak Website are more likely to be Converted.

Model Building:

- We created dummy variables for all categorical variables and we split the data into train and test 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 the data set to make the model more robust
- After building our first model we used the Variable inflation factor and p-values of the model to 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 top Feature that contributes to the decision are
  1. TotalVisits
  2. Total Time Spent on Website
  3. Lead Origin→Lead Add Form
  4. Lead Source→Welingak Website
  5. What is your current occupation→Unemployed
  6. What is your current occupation→Student

Learning:

- How to work in a team during a project
- How to handle data cleaning and preparations for a logistic regression.
- How to develop Logistic Regression model in python.
- How to create dummy variables on categorical columns
- How we can choose a cut off for model based on sensitivity and specificity.
- How to get list of variables from final model which contribute most towards the probability and help to solve business problem.