Problem Statement

An education company named X Education sells online courses to industry professionals who aim to optimize the conversion rate from potential customers.

Model Objective

Classification model, target is the ability that customers will buy the courses (0 or 1).

EDA Report

- There are missing values in various columns
- **Numerical columns**: 'Lead Number', 'Converted', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'
- **Target column**: Converted
- **Nominal variable**: 'Prospect ID', 'Lead Origin', 'Lead Source', 'Last Activity', 'Country', 'Specialization', 'How did you hear about X Education', 'What is your current occupation', 'What matters most to you in choosing a course', 'Magazine', 'Receive More Updates About Our Courses', 'Tags', 'Lead Quality', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'Lead Profile', 'City', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'I agree to pay the amount through cheque', 'Last Notable Activity'
- **Binaries variable:** 'Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'A free copy of Mastering The Interview'

Numerical data

So far there are nothing special except missing values

Categorical data

- Using chi square to estimate the correlation between categorical data and target columns. Among those values, some of them have no relationship:
  {'Search': 0.0},
  {'Magazine': 0},
  {'Newspaper Article': 0.0},
  {'X Education Forums': 0.0},

{'Newspaper': 0.0},
{'Receive More Updates About Our Courses': 0},
{'Update me on Supply Chain Content': 0},
{'Get updates on DM Content': 0},
{'I agree to pay the amount through cheque': 0}
So I decided to cut them off.

- Top three most related categorical features:
  {'Tags': 0.9335340637666211},
  {'Lead Quality': 0.6599449125917178}
  {'Last Activity': 0.3979550112712201}
- Remove column ID as well. This is my believe.

Conclusion by EDA

In conclusion, Most of the customers do not want to communicate by phone or email unless they are interested in the course and they are unemployed. In terms of numerical correlation, Total Time Spent on Website, Asymmetrique Profile Score and Asymmetrique Activity Score are features that are strongly linked with the conversion rate. On the other hand, categories are Tags, Lead Quality, Last Activity. Therefore, I propose the strategy: with unemployed customers, we should apply the call and email strategy to maximize the conversion rate. Otherwise, Apply the call strategy for customers that having tag: “Closed by Horizzon” because this category is the highest proportion in conversion rate.