Summary

Problem Statement:

- **X Education** wants to **increase** it's lead conversion from **30% to 80%**.
- The company wants to focus more on `**Hot Leads**`.

Objective:

- Design a model to assign lead scores based on the conversion likelihood.

Steps followed:

1. **Data Cleaning -**
   The data was pretty clean, but there were some null values and some bogus `Select` values. In the data cleaning step, the null values were dropped, and the `Select` values were replaced by NaN; and all the NaNs and the NAs or the null values were then replaced with `not provided`.
2. **EDA -**
   Relationships between variables was looked out for; a lot of categorical variables turned out to not be useful enough.
3. **Dummy Variables -**
   Categorical variables with high correlation were converted into dummies, out of which the `not provided` ones were discarded.

Findings

1. It was found that the variables that mattered the most in the potential buyers are (In descending order) :
    a. The total time spent on the Website.
    b. Total number of visits.
    c. When the lead source was:
        i. Google
        ii. Direct traffic
        iii. Organic search
        iv. Welingak website
    d. When the last activity was:
        i. SMS
        ii. Olark chat conversation
    e. When the lead origin is Lead add format.
    f. When their current occupation is as a working professional.

Keeping these in mind **X Education** can identify the `**Hot Leads**` and maximize their conversion rate.