Problem Description

An education company named X Education offers online courses for industry professionals. Many professionals visit their website daily, coming from various marketing sources like other websites and search engines such as Google. On the website, visitors might browse courses, fill out a contact form, or watch videos. When they fill out the contact form, providing their email address or phone number, they become leads. The company also gets leads from past referrals.

The sales team then contacts these leads via calls and emails. Typically, only about 30% of these leads end up purchasing a course, which is considered a poor conversion rate. For example, out of 100 leads acquired in a day, only around 30 are converted into customers.

To improve this conversion rate, X Education wants to identify the most promising leads, called 'Hot Leads.' If they can correctly identify these Hot Leads, the sales team can focus their efforts on these leads, increasing the chances of conversion.

To achieve this, the company plans to build a logistic regression model. This model will assign a lead score between 0 and 100 to each lead. A higher score indicates a higher likelihood that the lead will convert into a customer. This scoring system will help the sales team prioritize leads that are more likely to become customers, ultimately improving the lead conversion rate.

LEARNINGS

The biggest and the foremost learning is to handle the complex data. As it was the first time experience that we saw the select as the recorded value instead of null values and we have to consider the select values as NaN values and handle the data accordingly. As in the online survey the non compulsory question not answered are being recorded as select in the data.

Second learning was to understand the impact of the problem on the business so that the correct solution can be implemented to the problem.

Third learning was in Model selection. As we were confused between two models weather we should use decision tree formulation or logistic regression formulation for the problem and this lead to the new learning in implementing the model.