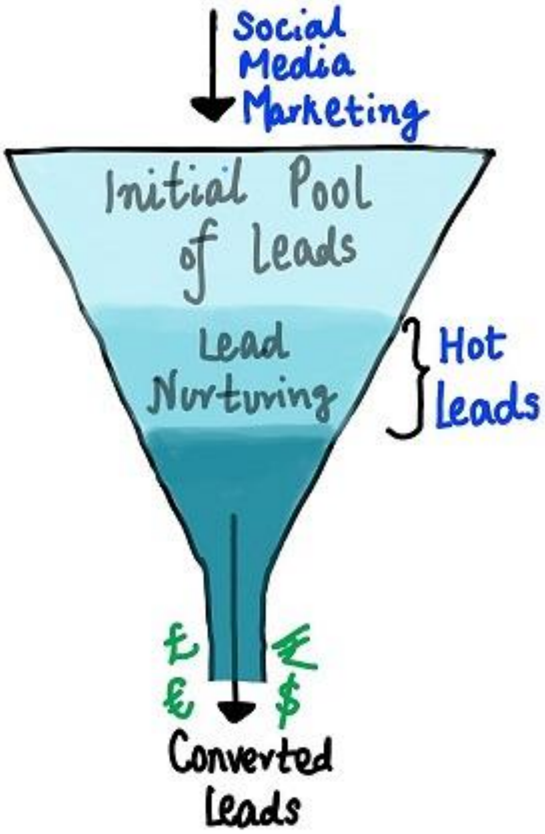
Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites, search engines, and even social media sometimes. Once these people land on the website, they might browse the courses, fill out a form for the course, or watch some videos. When these people fill out a form with their email address or phone number, they are classified as leads. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted into successful sales, while most of the leads do not. The typical lead to successful sale conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead-to-sale conversion rate is very poor. For example, if they acquire 100 leads in a day, only about 30 of them are converted into successful sales. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate would go up as the sales team would now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel.



Lead Conversion Process – Demonstrated as a Funnel

As you can see, there are a lot of leads generated in the initial stage (the initial pool of leads), but only a few of them come out as paying customers from the bottom (converted leads). In the middle stage (lead nurturing), you need to nurture the potential leads well (i.e., educate the leads about the product, constantly communicate, etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark estimate of the target lead conversion rate as being around 80%.

Data

You were given a leads dataset from the past that contained approximately 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on the Website, Total Visits, Last Activity, etc., which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’, which tells whether a past lead was converted or not, where 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out is the levels present in the categorical variables. Many of the categorical variables have a level called ‘Select’ which needs to be handled because it is as good as a null value (think why?).

Goals of the Case Study

There are quite a few goals for this case study. They are as follows:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads, which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., most likely to convert, whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company that your model should be able to adjust to if the company’s requirements change in the future, so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it out based on the logistic regression model you got in the first step. Also, make sure you include this in your final PowerPoint presentation, where you’ll make recommendations.

Results Expected

The following results are expected from this exercise:

1. A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
2. A word document filled with solutions to all the problems.
3. An overall approach of the analysis in a presentation, which should include the following:
    - Mention the problem statement and the analysis approach briefly.
    - Explain the results in business terms.
    - Include visualisations and summarise the most important results in the presentation.
4. A 500-word summary report explaining how you approached the assignment and the lessons you learnt.