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 and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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 while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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 should go up as the sales team will 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 (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating 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 of the target lead conversion rate to be around 80%.

Data

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on 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 wherein 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 are 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:

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. is 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 which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution:

The task was to improve the leads' conversion rates for X Education using logistic regression, the method was to go through a series of steps and learnings:

Assignment Approach:

1. Data Understanding and Preprocessing:

- Initially, I went through the entire dataset which contained several traits of leads such as season, weather data, and user interactions.

- Deal with missing values, encode categorical variables, and ensuring the numerical strength of the features by scaling was performed as data pre-processing.

2. Exploratory Data Analysis (EDA):

- EDA was used to establish the relationships between the variables and their impact on lead conversion.

- Detected some potential correlations and situations through histograms, scatter plots, and correlation matrices.

3. Model Building:

- Logistic regression was selected as the predictive model which could separate better between conversions and non-conversions because it is a binary classification model.

- I further distributed the data into training and testing sets in an 80-20 split manner so that I can assess the capacity of the model really efficient.

- The logistic regression model was trained using the training data and its performance was assessed using several metrics such as accuracy, precision, recall, and F1-score.

4. Model Evaluation and Interpretation:

- I created a confusion matrix to get the numbers of model's true positive, true negative, false positive, and false negative predictions to know how the model is.[Additional note: This may be due to the choosing of sentence lengths and some sentences that resemble initial text.]

- Coefficients were computed to find the variables that are most important for a lead to convert.

- Filtered out unnecessary features by using techniques like Recursive Feature Elimination (RFE), which made the model easier to understand and more efficient.

5. Business Recommendations:

- Suggested plans of action for X Education after the algorithm showed the following inputs, for instance, aggregating more possible customer leads in bursts and not calling the customers on the phone extremely when quotas are met quickly.

- Advancements in technology have made the shortening of communications and lead scoring systems common, thereby the optimization of lead management processes and efficiency.

Key Learnings:

- Data Preprocessing Importance: Clean and well-structured data are essential for precise model predictions. The common factors that cause variations in the model's performance are, the manner of dealing with missing data, the correct way of encoding categorical variables and ensuring that the temperatures are correctly normalized.

- Model Interpretability: The logistic regression model is easily explained by the coefficients that give us the key factors contributing to lead conversion.

- Feature Engineering and Selection: Reduced feature dimensionality helps to decrease memory consumption, rendering the process faster.

- Business Context Integration: The model's straight alignment with company needs and actionable strategies is crucial. Methods such as lead prioritization during peak seasons and automation during downtime are perfect instances of applying data-driven insights.

- Continuous Improvement: The iterative process of model refinement and its validation hold the key to success. Through techniques like cross-validation and regular performance monitoring, the models continue to be relevant in the changing business environment.