Lead Scoring Case Study- Summary

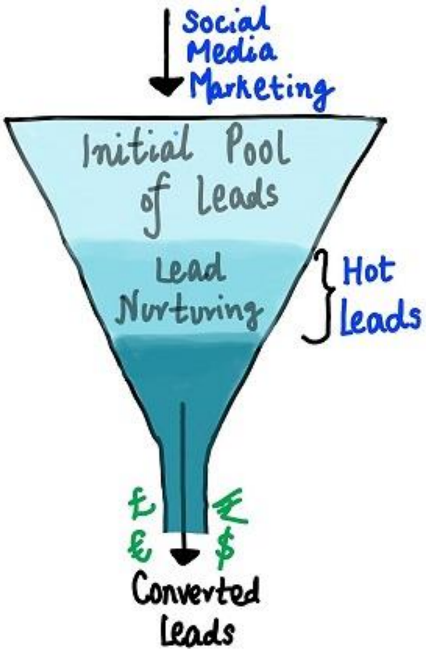
Problem Statement:

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 is now aimed in converting the initial pool of leads successfully as their customers by nurturing these hot leads.

Goals of the case study:

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.

Lead scoring helps us in assigning a score to every lead so that we can predict if that specific lead is going to convert or not and if yes, how probably is that lead going to convert. This is done by nurturing something called as hot leads as shown in the figure below.



The our main goal is to identify these potential hot leads which helps the sales team in educating them about the course, asking for one-on-one demo, referrals etc. Since the output of this process is a categorical type, building a predictive model using logistic regression is the optimal approach. With logistic regression, we can find the probabilities of the leads conversion with which we can assign a score to these leads and thus help sales team in lead conversion.

Steps followed in building the model:

1. Importing required libraries
2. Data inspection:
    - Checking the datatypes, shape, conversion rate data in the data frame.
3. Exploratory Data Analysis:

    - Data cleaning: Handling missing values and imputing missing values when very less.
    - Data visualization: Numerical and categorical
    - Handling outliers
4. Data preparation:
    - Creating dummy variables
    - Encoding necessary categorical variables
    - Handling data imbalance
    - Train test split of data set
5. Model building:
    - Feature Scaling
    - Using statsmodel and sklearn to build the model
    - Predicting based on latest model
    - Create confusion matrix and calculating the metrices
    - Find Optimal cutoff value
6. Conclusion on the model:
    - Model Conclusion
    - Comapring the metrices for train and test data set

The metrices for both the train and test turned out to be:

	Train data set	Test data set
Accuracy (%)	79.09	78.58
Sensitivity (%)	77.80	76.47
Specificity (%)	79.89	79.93
Positive predictive values (%)	70.48	71.01
Negative predictive values (%)	85.36	84.09

Top 3 variables which are contributing the most to the model

- What is your current occupation_Working Professional- This indicates that working professionals are more prone to be converting leads.
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website

By focusing on these aspects we can convert the hot leads to customers.